DANIEL
KAHNEMAN
PRIX NOBEL D'ÉCONOMIE
OLIVIER SIBONY
CASS R. SUNSTEIN

# NOISE

## Pourquoi nous faisons des erreurs de jugement et comment les éviter

Odile Jacob

Daniel Kahneman
PRIX NOBEL D'ÉCONOMIE

Olivier Sibony
Cass R. Sunstein

# NOISE

## Pourquoi nous faisons des erreurs de jugement et comment les éviter

*Traduit de l'anglais (États-Unis) par Christophe Jaquet*
*avec le concours d'Olivier Sibony*

Odile Jacob

Titre original : *Noise. A Flaw in Human Judgment*
Paru chez Little, Brown Spark (New York, Boston, London)

Copyright © 2021 by Daniel Kahneman, Olivier Sibony, and Cass R. Sunstein
All rights reserved

Pour la traduction française :

© ODILE JACOB, OCTOBRE 2021
15, RUE SOUFFLOT, 75005 PARIS

www.odilejacob.fr

ISBN : 978-2-7381-5706-5

Le code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5 et 3 a, d'une part, que les « copies ou reproductions strictement réservées à l'usage du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou réproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite » (art. L. 122-4). Cette représentation ou reproduction donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.

*Ce document numérique a été réalisé par Nord Compo.*

*Pour Noga, Ori et Gili – DK*
*Pour Fantin et Lélia – OS*
*Pour Samantha – CRS*

INTRODUCTION

# Deux types d’erreurs

Imaginons vingt personnes se donnant rendez-vous dans un stand de tir. Elles forment quatre équipes de cinq. Chaque équipe se partage une carabine. Chaque tireur ne tire qu’une fois. La figure 1 montre les résultats obtenus.

Dans un monde idéal, chaque tir atteindrait le centre de la cible.

C’est pratiquement le cas de l’équipe A. Ses tirs sont regroupés tout près du centre, formant un motif presque parfait.

De l’équipe B, nous dirons qu’elle souffre d’un *biais*, car ses tirs s’écartent tous du centre de la cible dans la même direction. La régularité du biais permet de formuler une prévision : si l’un des membres de l’équipe tirait de nouveau, nous parierions sans doute qu’il toucherait la cible à peu près au même endroit. La régularité du biais invite aussi à une explication causale : on soupçonne naturellement un problème sur la carabine qu’utilise cette équipe, par exemple une lunette de visée faussée.

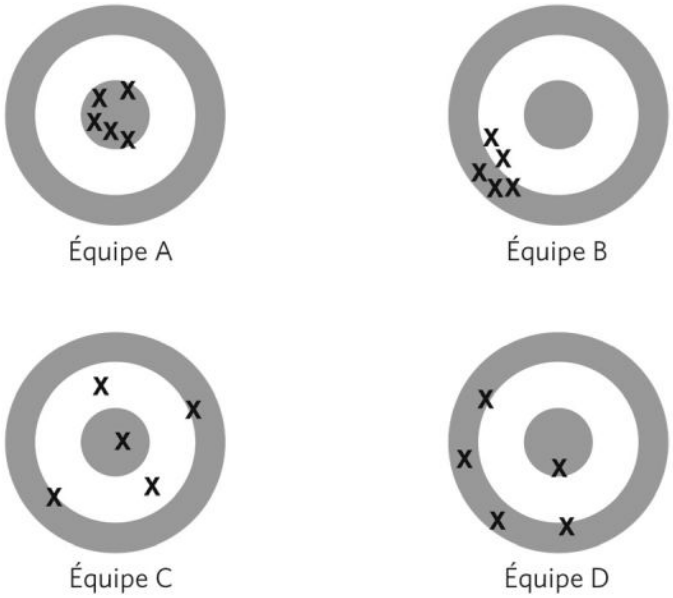



**Figure 1.** Quatre équipes au stand de tir.

Nous dirons de l'équipe C qu'elle est sujette au *bruit* (*noise* en anglais). Ses cinq tirs sont largement dispersés. Aucun biais n'est visible : le centre de gravité des impacts semble proche du mille. Si l'un des membres de l'équipe tirait de nouveau, nous serions bien en peine de prévoir quel point de la cible il aurait le plus de chances d'atteindre. En outre, aucune hypothèse ne vient à l'esprit qui puisse expliquer les résultats de cette équipe. Nous savons que ses membres sont des tireurs médiocres ; nous ignorons pourquoi ils sont sujets au bruit.

L'équipe D, enfin, est sujette tout à la fois au biais et au bruit. Comme l'équipe B, elle a tendance à toucher le quart sud-ouest de la cible ; et comme l'équipe C, ses tirs sont très dispersés.

Mais ce livre n'est pas un manuel de tir à la carabine. Son sujet est l'erreur humaine. Le biais et le bruit – l'écart systématique et la dispersion aléatoire – sont deux composantes différentes de l'erreur. C'est cette différence qu'illustrent nos quatre cibles[1].

Le stand de tir est une métaphore des erreurs qui peuvent entacher le jugement humain, et en particulier les décisions que prennent des individus pour le compte des organisations qui les emploient. Nous trouverons dans ces situations les deux types d'erreurs illustrés par la figure 1. Certains jugements souffrent de biais, ce qui veut dire qu'ils manquent généralement leur cible de la même manière. D'autres sont sujets au bruit, ce qui veut dire que des personnes qui devraient être d'accord arrivent à des conclusions très éloignées. Un grand nombre d'organisations, malheureusement, souffrent à la fois de biais et de bruit.

La figure 2 illustre une différence importante entre le biais et le bruit. Elle montre ce que nous observerions au stand de tir si nous ne voyions les cibles que de *dos*, sans savoir où se trouve leur centre.

On le voit, il n'est plus possible de dire laquelle de l'équipe A ou de l'équipe B est le plus près du centre de la cible. Mais un coup d'œil suffit pour dire que les équipes C et D sont sujettes au bruit, et pas les équipes A et B. Car la dispersion est ici tout aussi visible que sur la figure 1. Une des propriétés générales du bruit est qu'il est possible de l'identifier et de le mesurer sans rien savoir ni de la cible ni du biais.

Cette propriété générale est essentielle pour le propos de ce livre, car nos observations porteront souvent sur des jugements pour lesquels la bonne réponse n'est pas connue, et souvent même pas connaissable. Quand deux médecins posent deux diagnostics différents après avoir examiné un même patient, nous n'avons pas besoin de savoir ce dont souffre le malheureux pour étudier leur désaccord. Quand deux productrices de cinéma évaluent le marché potentiel d'un film, on peut mesurer l'écart entre leurs réponses sans savoir combien le film a finalement rapporté d'argent, ni même s'il a effectivement été produit. Pour mesurer le bruit, nous n'avons pas besoin de savoir où se trouve le centre de la cible : il suffit de la regarder de dos.

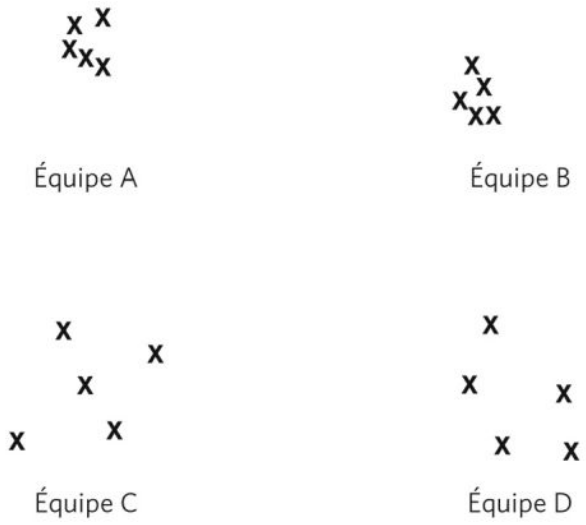



**Figure 2.** Les cibles vues de dos.

Si l’on veut comprendre l’erreur de jugement, il faut s’intéresser à la fois au biais et au bruit. Quelquefois, comme nous le verrons, le bruit est la composante la plus importante de l’erreur. Et pourtant, dans le débat public sur l’erreur humaine, et dans les organisations, partout dans le monde, l’existence du bruit est rarement reconnue. Le biais tient le premier rôle ; le bruit, quant à lui, est considéré au mieux comme un figurant, quand il n’est pas relégué dans la coulisse. Les biais ont fait l’objet de milliers d’articles et de centaines de livres, mais il est rare qu’on y évoque ne fût-ce qu’en passant la question du bruit. Il y a là un important déséquilibre. Le bruit est la face cachée de l’erreur. Ce livre va tenter de la mettre en lumière.

Le niveau de bruit qui affecte les décisions prises dans le monde réel est souvent scandaleusement élevé. Pour donner un avant-goût des données que nous passerons en revue dans ce livre, en voici quelques exemples préoccupants, dans des situations où la justesse des décisions a pourtant une grande importance :

- *La médecine*. Face au même patient, les médecins posent souvent un diagnostic différent, qu’il s’agisse du cancer de la peau ou des poumons,

d’une maladie du cœur, de la tuberculose, de la pneumonie, de la dépression ou de mille autres pathologies. Le bruit est particulièrement élevé en psychiatrie, domaine dans lequel le jugement subjectif joue évidemment un grand rôle. Mais il est aussi très présent dans des spécialités où l’on s’y attend le moins, comme la radiologie.

- *La protection de l’enfance*. Les travailleurs sociaux chargés de la protection de l’enfance doivent déterminer quels enfants sont en danger ou risquent d’être victimes de violences, et, quand c’est le cas, décider ou non de leur placement dans une famille d’accueil. Le système est sujet au bruit, car certains professionnels sont beaucoup plus enclins que d’autres à prendre une décision de placement. Des années plus tard, les enfants concernés seront plus nombreux à avoir des difficultés de vie : taux de délinquance plus élevé, grossesses précoces plus fréquentes, revenus inférieurs, etc.[2].

- *Les prévisions*. Qu’il s’agisse de prévoir les ventes d’un nouveau produit, le taux de chômage ou la probabilité d’une faillite d’entreprise, les prévisionnistes professionnels proposent des chiffres extrêmement variables. Non seulement ils ne sont pas d’accord entre eux, mais souvent ils ne le sont pas non plus avec eux-mêmes. On a ainsi demandé deux fois à un groupe de développeurs de logiciels, à quelques semaines d’intervalle, d’évaluer le temps nécessaire pour réaliser la même tâche : les durées estimées différaient de 71 % en moyenne[3].

- *Les demandes d’asile politique*. Aux États-Unis, la décision d’admettre ou non un demandeur d’asile s’apparente à une loterie. Une étude portant sur des affaires attribuées de manière aléatoire à différents juges montre que l’un d’eux donnait une réponse favorable à 5 % des requêtes qui lui étaient soumises, et un autre à 88 %. L’étude portait un titre

évocateur : *Refugee roulette*[4], « La roulette des réfugiés ». L'image de la roulette, ou de la loterie, reviendra souvent dans ce livre.

- *Les ressources humaines*. Les personnes qui font passer des entretiens d'embauche portent sur les mêmes candidats des appréciations très différentes. Une fois embauchés, les candidats n'en ont pas fini avec le bruit : leurs évaluations de performance sont elles aussi très variables et dépendent davantage de la personne qui procède à l'évaluation que de la performance évaluée.

- *La liberté sous caution*. Les juges américains doivent décider si un prévenu, dans l'attente de son procès, peut être remis en liberté sous caution ou s'il doit être placé en détention provisoire. Cette décision dépend largement de la personne du juge chargé de l'affaire. Certains sont bien plus cléments que d'autres avec l'ensemble des prévenus. Qui plus est, même des juges dont la propension moyenne à accorder une liberté sous caution est assez proche peuvent évaluer de façon très différente le risque de récidive ou de fuite d'un prévenu donné.

- *La police scientifique*. On pense souvent que les empreintes digitales sont un moyen infaillible d'identification. Or les experts ne sont pas toujours d'accord pour dire si les empreintes trouvées sur une scène de crime correspondent à celle d'un suspect. Et en plus de ne pas être d'accord entre eux, ils peuvent aussi se contredire et prendre des décisions différentes quand on leur présente des paires d'empreintes qu'ils ont déjà étudiées par le passé. Une variabilité similaire s'observe dans d'autres disciplines de la science médico-légale, y compris les analyses d'ADN.

- *Les brevets*. Les auteurs d'une étude faisant autorité ont identifié la présence de bruit dans l'attribution des brevets : « Que l'office des brevets accorde ou rejette un brevet est lié de façon significative au

hasard qui fait que le dossier sera confié à tel ou tel inspecteur », écrivent-ils[5]. Cette variabilité est évidemment problématique du point de vue de l'équité.

Ces exemples de bruit ne sont que la partie émergée d'un énorme iceberg. Chaque fois que l'on étudie le jugement humain, il y a de très fortes chances qu'on y trouve du bruit. Pour améliorer la qualité de nos jugements, nous devons nous attaquer au biais, mais aussi au bruit.

Ce livre se compose de six parties. Dans la première, nous explorons la différence entre le biais et le bruit, et nous montrons que les organisations peuvent être sujettes au bruit à un degré parfois ahurissant. Pour situer le problème, nous commençons notre tour d'horizon par deux domaines, relevant l'un du secteur public et l'autre du secteur privé : la justice pénale et l'assurance. Nous présentons dans les deux cas ce que nous appelons un audit de bruit : une étude qui mesure le niveau de désaccord entre des professionnels appelés à traiter de cas similaires au sein de la même organisation.

Dans la deuxième partie, nous examinerons la nature du jugement humain et verrons comment en mesurer l'exactitude – et l'erreur, qui se compose de biais et de bruit. Nous ferons apparaître une équivalence frappante dans les rôles respectifs de ces deux types d'erreurs. Nous découvrirons le bruit occasionnel : la variabilité dans les jugements que le même individu (ou le même groupe) produit sur la même question quand il y est confronté à plusieurs reprises. Nous verrons que le niveau de bruit occasionnel est particulièrement élevé dans les discussions de groupe, et que des facteurs qui devraient être insignifiants, comme l'ordre de prise de parole, en sont une source importante.

Les deux parties suivantes du livre sont essentiellement théoriques (les lecteurs qui s'intéressent principalement aux effets pratiques du bruit et à sa réduction pourront d'ailleurs sauter directement à la cinquième partie). La

troisième partie pose un regard plus approfondi sur un type de jugement qui a fait l'objet d'abondantes recherches : le jugement prédictif. Nous examinerons l'avantage décisif qu'ont les règles, les formules et les algorithmes sur les humains quand il s'agit de faire des prédictions : contrairement à une croyance répandue, cet avantage n'est pas tant dû à l'intelligence supérieure des règles qu'à l'absence de bruit qui les caractérise. Nous discuterons de la limite indépassable qui s'impose à la qualité de tout jugement prédictif – l'ignorance objective de l'avenir – et montrerons comment elle s'additionne au bruit pour limiter la qualité des prévisions.

La quatrième partie se tourne vers la psychologie du jugement. Nous y expliquons les principales causes du bruit. Parmi celles-ci, les différences interpersonnelles tiennent un rôle central. Elles peuvent être dues à une pondération différente des facteurs à prendre en compte, à une interprétation différente des échelles utilisées pour formuler un jugement, à la personnalité et au style cognitif des individus, et à bien d'autres raisons encore. Nous verrons pour quelles raisons le bruit est si souvent absent de notre analyse de l'erreur, et pourquoi il arrive si souvent que nous ne soyons pas surpris par des événements ou par des jugements que nous avions été absolument incapables de prévoir.

Dans la cinquième partie, nous aborderons une question essentielle d'un point de vue pratique : comment améliorer le jugement et prévenir l'erreur ? Pour ce faire, nous présenterons cinq études de cas dans des domaines connus pour être particulièrement sujets au bruit, et dans lesquels des efforts ont déjà été faits pour le réduire, avec des succès dont l'inégalité est très instructive : le diagnostic médical, l'évaluation de la performance, la science médico-légale, la décision d'embauche, la prévision en général. Ces études de cas nous donneront l'occasion de présenter un ensemble de techniques de réduction du bruit, réunies sous le nom d'« hygiène de la décision ». Nous conclurons cette partie en proposant une approche

structurée de l'évaluation des options, que nous appelons le *protocole d'évaluations intermédiaires* ; cette approche intègre plusieurs pratiques clés de l'hygiène de la décision et vise à produire moins de bruit et des jugements plus fiables.

Y a-t-il un « bon » niveau de bruit ? La sixième et dernière partie tentera de répondre à cette question. Peut-être le bon niveau n'est-il pas de zéro. Dans certains domaines, il est tout simplement impossible d'éliminer le bruit. Dans d'autres, cela coûterait trop cher. Dans d'autres encore, les efforts pour y parvenir pourraient porter atteinte à des valeurs fondamentales ou saper la motivation des individus en leur donnant le sentiment qu'ils ne sont que des rouages dans une machine. Quand les algorithmes font partie de la réponse, toutes sortes d'objections surgissent aussitôt ; nous répondrons à quelques-unes d'entre elles. Quoi qu'il en soit, le niveau actuel de bruit est inacceptable. Nous pensons qu'il est urgent que les organisations, dans le secteur privé comme dans la sphère publique, réalisent des audits de bruit pour mesurer le niveau de bruit dans leurs jugements et prennent des mesures énergiques pour le réduire. L'enjeu est non seulement de réduire les coûts considérables qu'engendrent les erreurs de jugement, mais aussi de remédier à des inéquités aussi fréquentes qu'invisibles.

C'est cette aspiration qui nous a conduits à formuler, en fin de chaque chapitre, des propositions pratiques présentées sous forme d'aphorismes. Vous pourrez les appliquer telles quelles ou les adapter aux questions qui vous tiennent à cœur, qu'il s'agisse de santé, de sécurité, de management, d'éducation, d'argent, d'emploi, ou de mille autres domaines dans lesquels nous exerçons quotidiennement notre jugement.

Comprendre le problème du bruit, et essayer de le résoudre, est un travail de tous les instants et une œuvre collective. Nous avons tous la possibilité d'y participer. Ce livre est écrit dans l'espoir que nous puissions nous en saisir.

PARTIE I

# Détecter le bruit

Est-il acceptable que deux personnes ayant des profils comparables et qui ont été convaincues du même crime écopent de peines radicalement différentes – par exemple, cinq ans de prison pour l'une et la liberté sous caution pour l'autre ? Hélas, cela arrive souvent. Certes, le système judiciaire est truffé de biais. Mais, dans le chapitre 1, c'est au bruit que nous nous intéressons, et en particulier à la croisade d'un juge réputé qui attira l'attention sur le phénomène, dénonça le scandale qu'il représentait et parvint à le réduire un peu (mais pas assez). L'histoire que nous allons raconter se passe aux États-Unis, mais le même problème se pose, et les mêmes solutions pourraient s'appliquer, dans beaucoup d'autres pays. Nous pensons même que dans certains d'entre eux, le problème du bruit est pire encore qu'aux États-Unis. Cet exemple démontre que le bruit peut être source de terribles injustices.

Pour frappant que soit l'exemple de la justice pénale, les enjeux peuvent être aussi très importants dans le secteur privé. Au chapitre 2, nous prenons l'exemple d'une grande compagnie d'assurances, où deux groupes de professionnels formulent des jugements importants : les souscripteurs, qui fixent les primes d'assurance des clients potentiels, et les experts en sinistres, qui estiment le coût des sinistres. On pourrait penser que ces

tâches sont simples et routinières, et s'attendre donc à ce que des professionnels compétents arrivent à peu près aux mêmes sommes. Nous avons réalisé une expérience soigneusement conçue – un audit de bruit – pour tester cette prédiction. Les résultats nous ont surpris, mais surtout, ils ont étonné et consterné la direction de l'entreprise, qui a constaté que le bruit lui coûte énormément d'argent. Ce cas illustre une autre conséquence néfaste du bruit : il peut provoquer des pertes économiques considérables.

Ces deux exemples impliquent des études réalisées sur un grand nombre de personnes, qui produisent elles-mêmes un grand nombre de jugements similaires. Or beaucoup de jugements importants ne sont pas récurrents mais *singuliers* : faut-il lancer un produit nouveau ? Devons-nous recruter quelqu'un qui n'a pas le profil habituel pour le poste ? Que faire quand se présente une opportunité d'affaire apparemment unique ? Et que décider quand survient une pandémie ? Il serait tentant de penser qu'il n'est pas possible de trouver du bruit dans des situations uniques comme celles-ci. Après tout, le bruit est une variabilité non désirée : comment pourrait-il y avoir de la variabilité quand la décision est singulière ? Nous répondrons à cette question dans le chapitre 3. Nous y montrerons que tout jugement, même dans une situation apparemment unique, est une possibilité parmi des myriades d'autres. Même un jugement singulier est sujet à beaucoup de bruit.

La conclusion qui se dégage de ces trois chapitres peut être résumée en une phrase, qui sera un des principaux leitmotive de ce livre : *dès lors qu'il y a jugement, il y a bruit, et plus qu'on ne l'imagine*. Essayons de l'évaluer.

CHAPITRE 1

# Crime et châtiment

Supposons que quelqu'un ait été reconnu coupable d'un crime – vol à l'étalage, possession d'héroïne, coups et blessures ou vol à main armée. À quelle peine sera-t-il condamné ?

La réponse ne devrait dépendre ni du juge à qui l'affaire a été confiée, ni du temps qu'il fait, ni du score de l'équipe de football locale lors du match de la veille. Il serait scandaleux que trois individus aux profils en tout point comparables et convaincus du même crime soient condamnés à des peines radicalement différentes : la liberté sous caution pour le premier, 2 ans de prison pour le deuxième et 10 ans pour le troisième.

Pourtant, ce scandale se produit quotidiennement.

Les juges, dans le monde entier, ont longtemps joui d'un pouvoir discrétionnaire dans le choix de la peine. Ce pouvoir discrétionnaire, considéré comme un gage de justice et d'humanité, a été célébré par les plus grands juristes. Ils estimaient que la peine criminelle devait être choisie sur la base d'un faisceau de facteurs : le crime lui-même, mais aussi la personnalité du prévenu et les circonstances. L'individualisation de la peine était un principe capital. Chaque prévenu avait le droit que sa situation soit examinée dans le plus grand détail. Tenter, au contraire, d'encadrer par des règles la décision des juges ne pouvait conduire qu'à une justice mécanique

et déshumanisée. Aux yeux de beaucoup, l'idée même de procès équitable semblait nécessiter un pouvoir discrétionnaire total du juge.

Cet enthousiasme universel commença à se tempérer dans les années 1970, pour une raison simple : un niveau assourdissant de bruit dans la décision judiciaire. En 1973, un juge célèbre, Marvin Frankel, attira l'attention de l'opinion sur le problème. Avant de devenir magistrat, Frankel avait été un défenseur fervent des droits de l'homme et de la liberté d'expression, et contribué à la création du Comité des avocats pour les droits humains, aujourd'hui connu sous le nom de Human Rights First. Et il pouvait être féroce.

S'agissant du bruit dans le système pénal, ce que découvrit Frankel le scandalisa. Il intitula son livre *Criminal Sentences : Law Without Order* [« Peines criminelles : la loi sans l'ordre »]. Comme il l'écrivit [1] :

> *Quand une personne était condamnée par un tribunal fédéral pour le braquage d'une banque, elle risquait une peine maximum de 25 ans de prison. Autrement dit, n'importe quel chiffre entre 0 et 25 ans. Je me suis vite rendu compte que le nombre d'années choisi dépendait moins de l'affaire ou de la personne du prévenu que de celle du juge, c'est-à-dire des opinions, des préférences et des biais du magistrat. Si bien que pour le même crime, le prévenu pouvait être condamné à des peines extrêmement différentes en fonction du juge à qui il avait affaire.*

Frankel ne s'appuyait sur aucune analyse statistique pour étayer sa thèse. Mais il présentait une série d'anecdotes éloquentes, qui montraient de fortes disparités dans le traitement de prévenus très semblables. Pour avoir contrefait des chèques d'un montant de 58,40 dollars pour le premier et 35,20 dollars pour le second, deux hommes, sans casier judiciaire, avaient été condamnés respectivement à 15 ans et à 30 *jours* de prison. Pour des faits similaires de détournement de fonds, un homme était condamné à 117

*jours* de prison, et un autre à 20 *ans*. Citant plusieurs autres exemples, Frankel dénonçait les « pouvoirs étendus et presque sans contrôle[2] » des juges fédéraux. Il y voyait un excès de pouvoir qui se traduisait par « la perpétuation quotidienne de cruautés arbitraires[3] », inacceptable dans « un État de droit qui dépend des lois, pas des hommes[4] ».

Frankel en appela au législateur pour mettre fin à cette « discrimination ». Par ce mot, il désignait la variabilité inexplicable des condamnations, c'est-à-dire, essentiellement, le bruit. Mais il s'inquiétait aussi de la présence de biais, c'est-à-dire de disparités raciales et socio-économiques. Pour combattre à la fois le bruit et les biais, il proposait que les différences de traitement des criminels ne soient autorisées que si elles pouvaient être « justifiées par des tests adaptés, élaborés et réalisés avec une objectivité suffisante pour garantir que les peines ne soient pas simplement les oukases personnels de quelques responsables publics, magistrats ou autres[5] ». Mieux encore, il appelait à une réduction du bruit au moyen « d'un profil détaillé ou d'une liste de facteurs comprenant, si possible, une forme de notation objective, par exemple chiffrée[6] ».

Écrivant au début des années 1970, Frankel n'allait pas jusqu'à prôner ce qu'il appelait le « remplacement des personnes par les machines ». Il n'en était cependant pas loin. Il était convaincu que l'État de droit « nécessitait un ensemble de règles impersonnelles, universellement applicables, liant autant les juges que tout un chacun ». Il défendait expressément l'utilisation de « l'informatique pour faciliter une réflexion ordonnée sur la peine[7] ». Il recommandait enfin la création d'une commission des peines[8].

Le livre de Frankel devint l'un des plus marquants dans l'histoire du droit pénal américain – et sans doute du droit pénal en général. Son réquisitoire manquait cependant de données chiffrées. Pour combler ce vide, plusieurs études examinèrent de façon méthodique le niveau de bruit inhérent à la justice pénale.

La première grande étude, réalisée en 1974, fut dirigée par Frankel lui-même. Il était demandé à cinquante magistrats fédéraux, issus de différents districts du pays, de prononcer une peine pour des prévenus hypothétiques dont les affaires étaient résumées dans un rapport identique. L'étude démontra que l'« absence de consensus était la norme[9] » et jugea la disparité des condamnations « stupéfiante[10] ». En fonction du juge, la peine d'emprisonnement variait de 1 à 10 ans pour un trafiquant d'héroïne[11] ; de 5 à 18 ans pour un braqueur de banque[12] ; et de 3 ans de prison (sans amende) à 20 ans de prison (et 65 000 dollars d'amende) pour une affaire d'extorsion de fonds[13]. Plus étonnant encore : dans seize cas sur vingt, les juges n'étaient pas unanimes pour prononcer ou non une peine d'emprisonnement.

Cette étude fut suivie de plusieurs autres, qui révélèrent toutes un niveau de bruit scandaleux. En 1977, par exemple, William Austin et Thomas Williams demandèrent à quarante-sept juges de juger les cinq mêmes affaires, chacune impliquant des délits relativement mineurs[14]. Les descriptifs des affaires comprenaient une synthèse des informations utilisées par les juges dans des cas réels de condamnation : l'accusation, les témoignages, le casier judiciaire du prévenu (s'il en avait un), et des éléments sur sa vie et sa personnalité. La principale conclusion se résumait en deux mots : « variation substantielle ». Dans une affaire de cambriolage, par exemple, la peine recommandée allait de 30 jours de prison (et 100 dollars d'amende) à 5 ans de prison. Dans une affaire de possession de marijuana, certains juges recommandaient la prison, d'autres la liberté sous caution.

Une étude beaucoup plus importante, réalisée en 1981, mobilisa deux cent huit juges fédéraux, confrontés aux seize mêmes affaires fictives[15]. Ses conclusions étaient stupéfiantes : « Il n'y a d'accord unanime sur le principe d'une peine d'emprisonnement que dans trois affaires sur seize. Et même quand la plupart des juges sont d'accord pour prononcer l'incarcération, sa

durée varie considérablement. Dans une affaire de fraude, la peine moyenne recommandée est de 8,5 ans, et la plus longue la perpétuité. Dans une autre, la peine moyenne est de 13 mois et la plus longue de 15 ans. »

Si révélatrices soient-elles, ces études, fondées sur des expériences contrôlées, sous-estiment très certainement l'ampleur du bruit dans la justice pénale. Dans le monde réel, les juges disposent d'informations autrement plus nombreuses et complètes que les descriptions schématiques fournies aux sujets de ces études. Bien sûr, certaines de ces informations supplémentaires sont intéressantes, mais comme nous le verrons, il existe de sérieuses raisons de penser que des informations non pertinentes, voire des facteurs mineurs et aléatoires, peuvent conduire les juges à des décisions extrêmement différentes. Ainsi, les juges ont tendance à accorder plus souvent la liberté sous caution en début de journée ou après une pause déjeuner que juste avant celle-ci [16]. Quand le juge a faim, il est plus sévère !

Une étude portant sur un millier de décisions rendues par des juges pour enfants montre que les juges prennent des décisions plus sévères le lundi (et aussi, mais dans une moindre mesure, le reste de la semaine) lorsque l'équipe de football locale a perdu son match du week-end [17]. Les prévenus noirs sont disproportionnellement victimes de ce surcroît de sévérité. Une autre étude, portant sur 1,5 million de décisions de justice rendues sur une période de trente ans, montre également que les juges sont plus sévères les jours qui suivent une défaite de l'équipe de football de leur ville que ceux qui suivent une victoire [18].

Une étude portant sur six millions de décisions rendues par des juges en France en l'espace de douze ans révèle qu'ils ont tendance à être plus cléments le jour de l'anniversaire du prévenu [19]. (On peut imaginer que les juges le sont aussi le jour de leur propre anniversaire, mais cette hypothèse n'a pas été testée.) Même la température extérieure peut influencer les juges [20]. Une étude portant sur 207 000 demandes d'asile politique, sur une période de quatre ans, fait apparaître un effet significatif des variations de la

température quotidienne : quand il fait chaud dehors, l'asile est accordé moins souvent. Si vous être persécuté dans votre pays et que vous demandez l'asile dans un autre, vous n'avez plus qu'à prier pour qu'il fasse frais le jour de votre audience.

## *Réduire le bruit dans les décisions judiciaires*

Dans les années 1970, les arguments du juge Frankel et les résultats empiriques qui les étayaient attirèrent l'attention d'Edward M. Kennedy, le frère de l'ancien président John F. Kennedy, qui comptait parmi les membres les plus influents du Sénat. Il en fut scandalisé et consterné. En 1975, il déposa une proposition de loi réformant la fixation des peines ; elle n'aboutit pas. Mais Kennedy était tenace. Données à l'appui, année après année, il revint soumettre son texte à ses collègues du Sénat. Et il finit par l'emporter. En 1984, pour remédier à la variabilité injustifiée des condamnations pénales, le Congrès adoptait le *Sentencing Reform Act* (loi sur la réforme de la peine).

La nouvelle loi entendait diminuer le bruit dans le système judiciaire en réduisant « le pouvoir discrétionnaire illimité que donne la loi aux juges et aux autorités compétentes en matière d'application des peines[21] ». Dénonçant « l'ampleur injustifiable » de la disparité des peines, le Congrès relevait notamment des études montrant que, dans la région de New York, les peines pour des affaires identiques pouvaient aller de 3 à 20 ans de prison[22]. Comme le recommandait le juge Frankel, la loi créait aussi la Commission des peines des États-Unis, dont la mission principale était claire : publier des « directives » obligatoires qui restreindraient l'éventail des peines applicables à chaque crime ou délit.

Ces directives furent adoptées par la Commission l'année suivante. Elles reposaient généralement sur les condamnations moyennes prononcées pour des crimes et délits similaires, à partir d'une étude portant sur dix mille affaires réelles. Soulignant les désaccords irréductibles au sein de la Commission, le futur juge à la Cour suprême Stephen Breyer, très impliqué dans le travail de la Commission, défendait l'utilisation de moyennes issues de la pratique passée : « Pourquoi la Commission n'a-t-elle pas pris son temps et tout remis à plat au lieu de se fonder simplement sur l'historique ? La réponse courte est : c'est impossible. C'est impossible, parce qu'il y a quantité d'excellents arguments dans tous les sens… Essayez de recenser tous les crimes existants et de les classer par ordre croissant de gravité… Regardez ensuite les réponses de vos amis à la même question, et voyez si elles sont identiques. Je peux vous dire que ce ne sera pas le cas[23]. »

Les directives imposent aux juges de prendre en compte deux facteurs avant de décider d'une peine : la nature de l'infraction et l'histoire criminelle du prévenu. À chaque infraction (crime ou délit) est attribué un niveau de gravité ; il y en a quarante-trois en tout. L'histoire criminelle du prévenu renvoie principalement au nombre et à la gravité de ses précédentes condamnations. Une fois ces deux facteurs pris en compte, les directives permettent un éventail relativement restreint de peines, la plus lourde ne pouvant dépasser la plus légère que de six mois ou de 25 %. Les juges ont le droit de s'écarter des directives pour tenir compte de circonstances aggravantes ou atténuantes, mais ils doivent alors le justifier devant une cour d'appel[24].

Bien qu'obligatoires, les directives ne sont pas rigides. Elles ne vont pas aussi loin dans la standardisation que l'aurait voulu le juge Frankel et laissent aux juges une marge de manœuvre significative. Plusieurs études, utilisant des méthodes différentes et portant sur des périodes différentes, arrivent cependant à la même conclusion : les directives diminuent le bruit.

Plus précisément, elles « réduisent la variation nette des peines [qui est] attribuable à l'aléa de l'identité du juge[25] ».

L'étude la plus aboutie sur l'impact des directives a été réalisée par la Commission elle-même[26]. Elle étudiait des affaires de braquage de banque, de distribution de cocaïne et d'héroïne et de détournement de fonds sur deux périodes : en 1985 (avant l'entrée en vigueur des directives), et en 1989-1990. Chaque affaire était comparée à d'autres, similaires au regard des critères pris en compte par les directives. Au sein de chaque groupe d'affaires comparables, les écarts entre juges étaient bien moindres pendant la seconde période.

D'après une autre étude, la différence moyenne entre juges dans la lourdeur des peines prononcées était de 17 %, ou 4,9 mois, en 1986-1987, et tombait à 11 %, ou 3,9 mois, en 1988-1993[27]. Une troisième étude, couvrant plusieurs périodes de temps, observait un succès similaire dans la diminution des disparités entre juges, définies comme les différences entre les peines moyennes prononcées par des juges pour des affaires similaires[28].

Malgré ces résultats conformes à leurs objectifs, les directives essuyèrent un feu roulant de critiques. De nombreux juges estimaient que les peines fixées par les directives pour certains crimes étaient trop sévères – une objection légitime, mais qui soulève une question de biais, pas de bruit. Plus intéressante pour nous était une autre objection, soulevée là encore par de nombreux magistrats : les directives, expliquaient-ils, étaient profondément injustes, parce qu'elles empêchaient les juges de tenir compte comme il se doit des circonstances particulières de chaque affaire. La réduction du bruit avait un prix : une automaticité inacceptable dans les décisions. Kate Stith, professeure de droit à Yale, et José Cabranes, juge fédéral, écrivaient ainsi : « Nous n'avons pas besoin de cécité mais de perspicacité et d'équité », ce qui « n'est possible que si le jugement prend en compte la complexité de chaque cas[29] ».

Cette ligne d'attaque devait exposer les directives à de vigoureuses contestations, tant juridiques que politiques. Ces contestations échouèrent toutes, jusqu'à ce que la Cour suprême, en 2005, pour des raisons techniques sans aucun rapport avec le débat résumé ici, décide d'invalider les directives[30]. Après la décision de la Cour, celles-ci perdirent leur caractère obligatoire et devinrent simplement indicatives. Il faut noter que les juges fédéraux s'en sont félicités : les trois quarts d'entre eux préféraient que les directives soient indicatives, contre 3 % seulement qui auraient préféré qu'elles restent obligatoires[31].

Quels furent les effets de la décision de la Cour suprême ? Crystal Yang, professeure de droit à Harvard, a examiné la question en analysant des quantités massives de données portant sur près de 400 000 condamnations réelles[32]. Sa conclusion principale est que les disparités entre juges, quelle que soit la façon dont on les mesure, ont augmenté de façon considérable à partir de 2005. Quand les directives étaient obligatoires, les prévenus jugés par un juge relativement sévère étaient condamnés à des peines plus longues de 2,8 mois, en moyenne, que ceux jugés par un juge moyen. Les directives devenues indicatives, cet écart *doubla*. Avec des accents qui évoquent ceux du juge Frankel quarante ans plus tôt, Yang écrit que ses « résultats posent de grands problèmes d'équité, car l'identité du juge désigné contribue de façon importante au traitement disparate de prévenus comparables condamnés pour des crimes identiques ».

L'étude de Yang le montre : libérés de la contrainte de directives devenues simplement indicatives, les juges se déterminent davantage en fonction des valeurs auxquelles ils sont personnellement attachés. Par exemple, les juges femmes ont tendance à user davantage de leur pouvoir discrétionnaire dans le sens de la clémence ; tout comme les juges nommés par des présidents démocrates. En outre, cette liberté ouvre la porte à certains biais en même temps qu'au bruit. Après la décision de la Cour suprême, on observe par exemple un accroissement significatif de la

disparité entre les peines prononcées pour les mêmes crimes contre des Afro-Américains et contre des Blancs.

Trois ans après le décès du juge Frankel en 2002, l'annulation des directives avait provoqué le retour de son pire cauchemar : la loi sans l'ordre.

---

L'histoire du combat du juge Frankel pour des directives en matière de justice pénale nous donne un remarquable aperçu des principaux points que nous allons examiner dans ce livre. Premièrement, *juger est un art difficile*, parce que le monde est compliqué et incertain. C'est une évidence dans le domaine judiciaire mais aussi dans la plupart des situations de jugement professionnel, comme ceux que produisent les médecins, les infirmières, les avocats, les ingénieurs, les enseignants, les architectes, les producteurs de cinéma, les cabinets de recrutement, les éditeurs, les responsables d'entreprise de toutes sortes, les sélectionneurs d'équipes de sport, et bien d'autres. Partout où il y a jugement, des désaccords sont inévitables.

Deuxièmement, *l'ampleur de ces désaccords est beaucoup plus grande qu'on ne l'imagine généralement*. S'il se trouve assez peu de personnes pour contester le principe d'individualisation des peines et le « pouvoir discrétionnaire » de le mettre en œuvre, presque tout le monde est choqué de découvrir l'ampleur des disparités qu'il engendre. La variabilité indue entre des jugements qui devraient idéalement être identiques – ce que nous appellerons ici le *bruit systémique* – peut provoquer des injustices, des erreurs de toutes sortes, et des coûts économiques élevés.

Troisièmement, *il est possible de réduire le bruit*. L'approche prônée par le juge Frankel et adoptée par la Commission des peines du Congrès – des règles et des directives – est une des nombreuses méthodes qui permettent de diminuer le bruit. Dans d'autres situations, d'autres méthodes conviennent mieux. Et certaines méthodes de réduction du bruit réduisent en même temps le biais.

Quatrièmement, *les efforts de réduction du bruit soulèvent souvent des objections et se heurtent à de grandes difficultés*. Il faut y répondre aussi, ou le combat contre le bruit sera voué à l'échec.

### À propos du bruit dans les décisions judiciaires

« Les expériences montrent des disparités bien plus grandes qu'on ne l'imagine entre les peines recommandées pour des affaires identiques. Cela ne peut pas être juste. »

« La peine ne devrait pas dépendre de l'humeur du magistrat pendant l'audience ou de la température extérieure. »

« Les directives sont un des moyens de limiter le bruit. Mais beaucoup de gens s'y opposent, parce qu'elles limitent le pouvoir discrétionnaire du juge, qu'ils estiment nécessaire pour assurer l'individualisation de la peine. »

« Pourtant, ces objections sont suspectes. Bien sûr, chaque cas est unique et chaque prévenu doit être traité individuellement. Mais idéalement, la peine ne devrait pas dépendre du juge auquel est confiée l'affaire. »

CHAPITRE 2

# Un bruit systémique

C’est dans un décor bien moins impressionnant que celui d’une cour d’assises que nous avons pour la première fois rencontré le bruit. Deux d’entre nous avaient en effet été engagés comme consultants par une grande compagnie d’assurances. Nous nous doutons, bien sûr, que tous nos lecteurs ne sont pas des passionnés d’assurances. Mais ce que nous avons découvert à cette occasion montre que le bruit peut coûter beaucoup d’argent à une société privée, permet de comprendre pourquoi ce problème passe souvent sous les radars, et illustre ce qu’il est possible de faire pour le résoudre.

Les dirigeants de la compagnie d’assurances avaient décidé de se pencher sur les jugements importants qui engagent l’entreprise. Ils s’interrogeaient sur l’opportunité de prendre des mesures pour faire en sorte que ces jugements soient plus cohérents – c’est-à-dire pour en réduire le bruit. Deux choses au moins leur semblaient claires. D’une part, la cohérence est souhaitable : idéalement, deux jugements sur la même question devraient être identiques. D’autre part, une cohérence parfaite n’est pas possible, puisque ces jugements sont informels et en partie subjectifs. Un peu de bruit est inévitable.

Là où les dirigeants de la compagnie n’étaient pas d’accord, c’est sur l’ampleur du bruit. La plupart d’entre eux ne pensaient pas qu’il puisse

représenter un problème significatif pour leur entreprise. C’est ainsi qu’ils prirent la décision louable de mener une expérience : ce que nous appelons un *audit de bruit*. Le résultat s’avéra surprenant, et nous fournit une parfaite illustration du problème du bruit.

## *Une loterie qui fait du bruit*

Toute grande entreprise emploie de nombreux professionnels autorisés à produire des jugements qui engagent la société. Ainsi, dans la compagnie d’assurances qui nous occupe, un nombre considérable de souscripteurs sont chargés de fixer le niveau des primes d’assurance. Si, par exemple, une banque demande à la compagnie de l’assurer contre les pertes dues à des fraudes commises par ses traders, un souscripteur doit évaluer ce risque et proposer la prime d’assurance correspondante. La compagnie emploie aussi des experts en sinistres, qui estiment le coût des sinistres déclarés par les assurés et qui négocient avec eux en cas de litige sur leur indemnisation.

Toutes les grandes agences de la compagnie comptent plusieurs souscripteurs qualifiés dans un domaine donné. Quand une demande d’estimation arrive, elle est confiée à la première personne qui se trouve être disponible. De fait, le choix du souscripteur qui la réalisera relève de la loterie.

Les estimations des souscripteurs ont des conséquences importantes pour la compagnie. Une prime élevée est avantageuse si elle est acceptée, mais le client potentiel risque de trouver l’addition trop salée et de s’adresser à la concurrence. Une prime sous-évaluée aura plus de chances d’être acceptée, mais elle est naturellement moins avantageuse pour la compagnie. Pour chaque risque, il existe donc un « juste prix », un montant idéal qui n’est ni trop élevé ni trop faible. Il y a de bonnes chances que le

jugement moyen d’un groupe important de professionnels n’en soit pas très éloigné. Mais les estimations supérieures ou inférieures au juste prix sont coûteuses : c’est là que le bruit, c’est-à-dire la variabilité des jugements, affecte les résultats de l’entreprise.

La mission des experts en sinistres est également très importante pour les finances d’une compagnie d’assurances. Supposons par exemple qu’une déclaration de sinistre soit faite au nom d’un ouvrier qui a perdu définitivement l’usage de sa main droite dans un accident du travail. Le sinistre est confié à un des nombreux experts qualifiés, choisi, comme dans le cas des souscripteurs, parce qu’il se trouve être disponible. L’expert rassemble les faits et fournit une estimation du coût final du sinistre pour la compagnie. C’est aussi à l’expert qu’il reviendra de négocier avec le demandeur ou avec son avocat pour s’assurer qu’il recevra l’indemnité promise dans la police d’assurance, tout en veillant à éviter que la compagnie ne paie plus qu’il ne faut.

L’estimation initiale de l’expert est importante, parce qu’elle fixe un objectif implicite dans les négociations futures avec le demandeur. La compagnie d’assurances est par ailleurs légalement tenue de provisionner le coût estimé de chaque sinistre (c’est-à-dire de mettre de côté dans ses comptes de quoi le payer). Là encore, il existe une valeur idéale du point de vue de la compagnie. L’avocat du demandeur pourra porter l’affaire devant un tribunal si la proposition faite par la compagnie lui paraît insuffisante. D’un autre côté, une provision trop généreuse peut donner à l’expert une latitude trop grande et l’amener à accepter des demandes excessives. Le jugement de l’expert est donc important pour la compagnie, et plus encore pour l’assuré.

Notre choix du mot « loterie » souligne ici le rôle du hasard dans le choix du souscripteur ou celui de l’expert à qui une affaire donnée est confiée. En temps normal, seul ce professionnel produit un jugement, et

personne ne peut savoir ce qui se serait passé si quelqu'un d'autre avait été choisi à sa place.

Nous ne sommes pas hostiles par principe aux loteries. Elles ont leur utilité quand il s'agit d'allouer des « biens », par exemple des cours à l'université, ou des « maux », comme la conscription militaire. Mais les loteries en matière de jugement, celles dont nous parlons ici, n'allouent rien du tout. Elles produisent seulement de l'incertitude. Imaginons un instant une compagnie d'assurances idéale, dans laquelle tous les souscripteurs fixeraient chaque prime au niveau optimal du « juste prix » – donc avec zéro bruit – mais où une loterie viendrait ensuite modifier de manière totalement aléatoire la proposition faite au client. Il s'agirait manifestement là d'un système absurde, où la loterie n'aurait aucune justification. Un système dans lequel le résultat dépend très largement de la personne désignée par le sort pour produire un jugement professionnel n'a aucune justification non plus.

## *L'audit de bruit, révélateur de bruit systémique*

La loterie qui désigne un tireur pour représenter son équipe ou celle qui désigne une juge pour prononcer une peine pénale créent de la variabilité. Mais, en temps normal, cette variabilité reste invisible, puisque aucune comparaison n'est possible. L'audit de bruit est un moyen de la rendre visible. Les études comparant les peines prononcées par plusieurs juges fédéraux dans le même dossier étaient des exemples d'audits de bruit. Dans un audit de bruit, le même problème est traité par plusieurs individus, et la variabilité de leurs réponses éclate au grand jour.

Les jugements des souscripteurs et des experts en sinistres se prêtent particulièrement bien à cet exercice, car leurs décisions reposent sur des informations écrites. Pour préparer l'audit de bruit, nous avons élaboré un descriptif détaillé de cinq affaires représentatives pour chacun des deux groupes (les souscripteurs et les experts en sinistres). Nous avons ensuite demandé à ces professionnels d'en évaluer deux ou trois chacun, de façon indépendante, sans leur dire que le but de l'étude était d'examiner la variabilité de leurs jugements[1].

Avant de poursuivre la lecture, nous vous suggérons de réfléchir un instant à la question suivante : dans une compagnie d'assurances bien gérée, si l'on choisit au hasard deux souscriptrices qualifiées ou deux expertes en sinistres compétentes, et si on leur confie exactement la même affaire, de combien leurs deux estimations divergeront-elles ? Ou, plus précisément : de combien sera la différence entre les deux estimations, exprimée en pourcentage de leur moyenne ?

Nous avons posé cette question à de nombreux cadres de la compagnie, puis, dans les années suivantes, à un grand nombre de dirigeants d'entreprise. Étonnamment, une réponse revient beaucoup plus souvent que les autres. La réponse la plus fréquente chez les cadres de la compagnie d'assurances était 10 %. Lorsque nous avons interrogé huit cent vingt-huit P-DG et cadres supérieurs de différents secteurs d'activité, 10 % était à la fois la réponse médiane et la plus fréquente (la deuxième était 15 %). Une différence de 10 % signifie, par exemple, qu'un des deux souscripteurs a fixé une prime d'assurance de 9 500 dollars et l'autre de 10 500 dollars. C'est une différence non négligeable, mais dont on peut en effet penser qu'il n'est pas impossible pour une entreprise de la tolérer.

Malheureusement, notre audit de bruit a révélé des différences bien plus importantes. L'écart médian entre deux estimations de la même prime d'assurance était de 55 % de leur moyenne : *cinq fois plus* que ce à quoi s'attendaient la plupart des gens, y compris les cadres dirigeants de la

compagnie. Ce résultat signifie, par exemple, que si la première souscriptrice estime une prime à 9 500 dollars, la seconde ne l'estime pas à 10 500 mais à 16 700 dollars ! Quant aux experts en sinistres, la différence médiane entre deux de leurs estimations était de 43 %. Encore faut-il souligner que ces résultats sont des médianes : dans la moitié des cas, donc, la différence entre deux jugements était encore plus importante.

Les dirigeants de la compagnie d'assurances ont vite compris que ce niveau de bruit posait un problème coûteux à leur entreprise. L'un d'entre eux a même estimé que le coût annuel du bruit dans la fixation des primes d'assurance – comprenant à la fois les clients perdus à cause d'estimations excessives et les pertes financières dues aux primes sous-estimées – se comptait en centaines de millions de dollars.

Quelle était, dans chaque cas, l'erreur ? Qu'on parle de l'erreur moyenne ou de celle de chaque professionnel, nul ne pouvait le dire précisément. Personne, en effet, ne pouvait déterminer avec précision le « juste prix » des primes ni celui des sinistres. C'est bien pour cela que le jugement des professionnels est indispensable ! Mais souvenez-vous du stand de tir : il n'est pas nécessaire de savoir où se trouve le centre de la cible pour mesurer la dispersion des points d'impact au dos de celle-ci. De même, il n'est pas nécessaire de connaître le juste prix pour mesurer le bruit dans les évaluations – et pour comprendre que ce bruit pose un problème de taille. L'audit de bruit montrait clairement que le sort d'un client dépendait très largement de la loterie désignant l'employé chargé de son dossier. Les clients n'auraient sans doute pas été très heureux d'apprendre qu'on les faisait participer à cette loterie sans leur consentement. De façon plus générale, quand vous avez affaire à une organisation, vous comptez sur elle pour produire des jugements fiables et cohérents. Vous ne vous attendez pas à y trouver du bruit systémique.

# *Variabilité indésirable ou diversité souhaitable*

L’une des caractéristiques du bruit *systémique* est qu’il est *indésirable*. Et avant de poursuivre, nous nous devons bien sûr de rappeler que la variabilité des jugements n’est pas *toujours* indésirable.

Prenons les questions de préférence ou de goût. Quand dix critiques de cinéma regardent le même film, quand dix œnologues goûtent le même vin, ou quand dix lecteurs commentent le même roman, nous ne nous attendons pas à ce qu’ils aient la même opinion, bien au contraire. La diversité des goûts est chose normale et même souhaitable : personne ne voudrait vivre dans un monde où tout le monde aurait exactement les mêmes préférences et les mêmes aversions (enfin, presque personne). Là où la diversité des goûts peut contribuer à des erreurs, c’est quand un individu prend son goût personnel pour un jugement qualifié. Si, par exemple, un producteur de cinéma passionné par l’histoire du téléphone à cadran décide d’approuver un scénario de long métrage sur ce sujet, et si son film est un échec commercial parce que le sujet n’intéresse absolument personne d’autre, son jugement professionnel aura été pris en défaut.

La variabilité des jugements est également normale et appréciable dans les environnements concurrentiels, où les bons jugements gagnent (et les mauvais perdent). Quand plusieurs entreprises concurrentes, voire plusieurs équipes au sein de la même entreprise, rivalisent pour apporter des solutions innovantes au même besoin du marché, il n’est pas forcément souhaitable qu’elles adoptent la même approche. Il en va de même quand plusieurs équipes de chercheurs s’attaquent à un défi scientifique, comme le développement d’un vaccin : il est alors largement préférable qu’elles l’envisagent sous des angles différents. Même les prévisionnistes se comportent quelquefois comme des joueurs en situation de concurrence. L’analyste qui est seul à avoir anticipé une récession imprévue est sûr de se

faire une réputation, tandis que celui qui va toujours dans le sens du consensus ne sortira jamais de l'anonymat. Dans toutes ces situations, la variabilité des idées et des jugements est bienvenue, pour une raison distincte : le jugement n'y constitue qu'une première étape. Dans une seconde étape, les résultats des jugements seront comparés entre eux, et le meilleur l'emportera. Sur un marché comme dans la nature, la sélection ne peut pas fonctionner sans variation.

Les questions de goût et les environnements concurrentiels posent des problèmes intéressants. Mais les jugements à propos desquels nous parlons de bruit sont ceux dont la variabilité est indésirable. Quand nous parlons de bruit systémique, nous décrivons le fonctionnement de « systèmes », qui ne sont pas des marchés, mais des organisations. Pour illustrer cette distinction, considérons l'exemple d'une action dont la valeur est évaluée différemment par différents traders : certains gagneront de l'argent, d'autres en perdront. Il faut, dit-on, des désaccords pour faire un marché. Mais si ces deux traders travaillent dans la même entreprise, si l'un d'eux est choisi au hasard pour faire cette évaluation, et si l'entreprise suit sa recommandation alors que sa collègue aurait produit une évaluation très différente, alors l'entreprise souffre de bruit systémique, et c'est un problème.

Cette observation est sensiblement celle qu'ont faite les dirigeants d'une grande société de gestion d'actifs, quand, après avoir entendu parler de nos premiers résultats, ils entreprirent de réaliser leur propre audit de bruit. Pour ce faire, ils demandèrent à quarante-deux investisseurs expérimentés travaillant dans la société d'estimer le « juste prix » d'une action (c'est-à-dire le prix de marché auquel ils ne trouveraient justifié ni de la vendre ni de l'acheter). L'estimation se fondait sur un descriptif de l'entreprise concernée, comprenant les comptes de résultat, les bilans et les flux de trésorerie des trois années antérieures, ainsi que les projections pour les deux années à venir. Le bruit médian, mesuré de la même manière que dans la compagnie d'assurances, était de 41 %. Des différences aussi grandes

entre les analystes d'une même entreprise, formés à utiliser les mêmes méthodes d'évaluation, ne sont pas bon signe.

Chaque fois qu'une personne est choisie au hasard pour produire un jugement au sein d'un ensemble d'individus ayant les mêmes qualifications, le bruit est un problème. C'était le cas pour cette société de gestion d'actifs, pour la justice pénale et pour la compagnie d'assurances. C'est aussi le cas de nombreuses organisations : en pratique, c'est souvent le hasard qui décide quel médecin vous verra à l'hôpital, quel inspecteur évaluera votre brevet, ou quel employé du service client traitera votre réclamation.

Une opinion répandue veut que cette variabilité n'ait pas beaucoup d'importance : ne dit-on pas que « les erreurs aléatoires s'annulent les unes les autres » ? Il y a là un raccourci dangereux. Certes, les erreurs de jugement positives et négatives ont tendance à s'annuler – quand elles concernent *la même affaire*. Nous verrons d'ailleurs comment cette propriété peut être mise à contribution pour réduire le bruit. Mais le bruit systémique ne concerne pas des jugements multiples de la même affaire : il provient de jugements variables sur *des affaires différentes*. Quand une police d'assurance est trop chère et une autre trop bon marché, rien ne sert de prétendre que les prix sont « en moyenne » corrects : la compagnie d'assurances a commis deux erreurs coûteuses. Et quand deux criminels qui auraient dû être tous deux condamnés à 5 ans de prison écopent l'un de 3 ans et l'autre de 7 ans, justice n'a pas été rendue « en moyenne ». Dans les systèmes sujets au bruit, les erreurs ne s'annulent pas : elles se cumulent.

## *L'illusion d'accord*

Il existe une littérature scientifique abondante sur le bruit en matière de jugement professionnel. Comme nous connaissions cette littérature, les

résultats de l’audit de bruit dans la compagnie d’assurances n’étaient pas une grande surprise pour nous. Ce qui nous a étonnés, en revanche, ce sont les réactions des dirigeants de la compagnie lorsque nous leur avons fait part de nos résultats : personne dans la compagnie ne s’attendait à un pareil niveau de bruit. Aucun des dirigeants n’a émis de doutes sur la validité de l’audit ni prétendu que le niveau de bruit mis en évidence était tolérable ; mais le problème du bruit et son coût considérable étaient une révélation. Le bruit était l’équivalent d’une fuite d’eau dans la cave : s’il était toléré, ce n’était pas parce qu’il était considéré comme acceptable, mais parce que personne ne l’avait détecté.

Comment cela était-il possible ? Comment des professionnels ayant la même mission, partageant les mêmes bureaux, pouvaient-ils diverger autant les uns des autres sans s’en rendre compte ? Comment leurs dirigeants étaient-ils passés à côté de cette réalité, dont ils comprenaient pourtant immédiatement qu’elle représentait une menace grave pour la performance et la réputation de leur entreprise ?

Il devint vite apparent que le problème du bruit systémique reste généralement invisible dans les organisations, et que cette inattention au bruit est, en elle-même, un phénomène aussi intéressant que sa prévalence. Ce que les audits de bruit nous montraient, c’est que des professionnels respectés – et les organisations qui les emploient – parvenaient à entretenir une *illusion d’accord*, alors qu’ils étaient en réalité en désaccord profond dans leurs jugements professionnels quotidiens.

Pour comprendre l’illusion d’accord, imaginez-vous à la place d’un souscripteur dans une compagnie d’assurances, par une journée de travail ordinaire. Vous avez plus de cinq ans d’expérience, vous savez que vous êtes respecté par vos collègues, et ce respect est réciproque. Vous savez, en somme, que vous êtes bon dans votre travail. Aujourd’hui, vous avez procédé à une analyse approfondie d’un ensemble de risques complexes auxquels est exposée une banque, et vous avez conclu qu’une prime de

200 000 dollars est appropriée. Le problème est certes délicat, mais pas beaucoup plus que ceux que vous traitez chaque jour de la semaine.

Imaginez maintenant que vos collègues de bureau aient reçu les mêmes informations que vous et procédé à une estimation du même risque. Le croiriez-vous si l'on vous disait qu'au moins la moitié d'entre eux ont proposé une prime supérieure à 255 000 dollars ou inférieure à 145 000 dollars ? L'idée est difficile à avaler. Il y a d'ailleurs fort à parier que les souscripteurs qui ont appris les résultats de l'audit de bruit, même s'ils en acceptaient la validité, n'ont jamais vraiment cru que ses conclusions s'appliquaient à leur cas personnel.

De manière générale – et pas seulement dans les jugements professionnels – la plupart d'entre nous pensons généralement que « le monde est comme il est parce que c'est comme ça ». À partir de là, il n'y a qu'un pas à faire pour croire aussi que « les autres voient le monde à peu près comme je le vois ». Ces croyances, qui définissent le *réalisme naïf*, sont essentielles à l'idée que nous nous faisons d'une réalité partagée avec autrui. Nous les remettons rarement en question[2]. Nous n'avons qu'une seule interprétation du monde, et nous ne faisons généralement pas beaucoup d'efforts pour en chercher d'autres. Notre interprétation nous suffit, parce que nous la vivons comme vraie. Personne ne passe son temps à imaginer d'autres façons de se représenter la réalité qu'il voit autour de lui.

S'agissant de jugements professionnels, l'idée que les autres voient le monde à peu près comme nous est confortée chaque jour de multiples façons. D'abord, nous partageons avec nos collègues un langage commun, ainsi qu'un ensemble de règles qui définissent les facteurs à prendre en compte dans nos décisions. Nous faisons parfois l'expérience rassurante d'être en accord avec nos collègues pour condamner les jugements qui violent clairement ces règles. Si nous avons des désaccords occasionnels avec certains collègues, nous les percevons comme un manque de jugement

inhabituel de leur part. Tout cela nous permet de ne pas remarquer que les règles sur lesquelles nous sommes d'accord sont en réalité vagues : suffisantes, sans doute, pour éliminer certaines possibilités, mais loin d'être assez précises pour spécifier la réponse à chaque cas particulier. Nous pouvons vivre confortablement avec nos collègues sans jamais nous rendre compte qu'en réalité, ils ne voient pas le monde de la même manière que nous.

Nous avons eu l'occasion de demander à une responsable expérimentée dans le département souscription d'une compagnie d'assurances de nous parler de son expérience professionnelle. « Quand j'étais une petite nouvelle, nous a-t-elle expliqué, je discutais des trois quarts des affaires avec mon supérieur […]. Après quelques années, je n'en avais plus besoin – et aujourd'hui, je suis considérée comme une experte. […] Avec le temps, je suis devenue de plus en plus sûre de mon jugement. » Comme un grand nombre d'entre nous, cette personne avait pris confiance en son jugement d'une seule manière : en l'exerçant.

La psychologie de ce processus est connue. La confiance se nourrit de l'expérience subjective que nous avons lorsque nous produisons des jugements avec une aisance et une facilité croissantes, notamment parce qu'ils ressemblent aux jugements que nous avons déjà produits dans des cas similaires. Au fil du temps, cette assureuse avait appris à être d'accord avec celle qu'elle avait été dans le passé, et sa confiance dans ses propres jugements s'était accrue. Rien ne suggère – du moins après la phase initiale d'apprentissage – qu'elle avait appris à être d'accord avec les autres, qu'elle avait vérifié dans quelle mesure elle l'était, ni même qu'elle avait fait des efforts pour que sa pratique ne s'écarte pas trop de celle de ses collègues.

Dans le cas de la compagnie d'assurances, seul l'audit de bruit a pu faire voler en éclats l'illusion d'accord. Mais si le réalisme naïf explique en partie que les professionnels n'aient pas conscience du bruit dans leurs

jugements, comment se fait-il que les dirigeants de la compagnie, de leur côté, aient si longtemps ignoré le problème ?

Plusieurs explications sont ici possibles, mais celle qui semble jouer un rôle majeur dans de nombreux environnements professionnels est simplement l'inconfort du désaccord. La plupart des organisations préfèrent le consensus et l'harmonie au désaccord et au conflit. Les procédures en place semblent souvent conçues pour minimiser la fréquence d'exposition à des désaccords et, quand ces désaccords surviennent, pour leur trouver une explication et y mettre ainsi fin.

Nathan Kuncel, un professeur de psychologie à l'Université du Minnesota réputé pour ses travaux sur la prédiction de la performance, nous a confié une anecdote qui illustre très bien ce problème. Kuncel aidait le service des admissions d'une université à revoir son processus de sélection. Le processus en vigueur exigeait qu'une première personne lise chaque dossier de candidature, lui attribue une note, puis remette le dossier ainsi noté à une deuxième personne, qui le lisait et le notait à son tour. Kuncel suggéra – pour des raisons qui vont devenir évidentes au fil de ces pages – de cacher la première note au second évaluateur, afin de ne pas l'influencer. Que lui répondit-on ? « C'est ce que nous faisions avant, mais cela provoquait tellement de désaccords que nous sommes passés au système actuel. » Dans bien des organisations, éviter les conflits est un objectif au moins aussi important que prendre de bonnes décisions.

Examinons un autre outil utilisé par beaucoup d'entreprises : l'analyse *a posteriori* des jugements malheureux. Comme mécanisme d'apprentissage, ces « autopsies », ou « retours d'expérience », peuvent être très utiles. Mais si une erreur a vraiment été commise, au sens où un jugement s'est écarté des normes en vigueur dans la profession, il ne sera pas très compliqué d'en convenir. On conclura sans difficulté que le jugement erroné était éloigné du consensus des experts. Il sera d'ailleurs tentant, par la même occasion, de considérer cet écart comme l'erreur isolée d'une brebis galeuse : les

mauvais jugements sont beaucoup plus faciles à identifier que les bons. Pourtant, dénoncer les erreurs grossières et montrer du doigt ceux qui les ont commises n'aide guère les professionnels à prendre conscience d'une autre réalité : quand ils produisent des jugements que personne ne vient remettre en question, ils sont bien plus souvent en désaccord qu'ils ne le pensent. Paradoxalement, la facilité avec laquelle ils trouvent un consensus sur les erreurs de jugement peut même renforcer leur illusion d'accord. Et les empêcher de tirer la véritable leçon : l'omniprésence du bruit systémique.

---

Nous espérons que vous commencez à partager notre thèse : le bruit systémique est un problème sérieux. Son existence n'est pas une surprise, car le bruit est une conséquence de la nature informelle du jugement. Cependant, comme nous le verrons dans ce livre, chaque fois qu'une organisation se penche sérieusement sur la question, le niveau de bruit qu'elle observe est un choc.

Notre conclusion est donc simple : dès lors qu'il y a jugement, il y a bruit – et plus qu'on ne l'imagine.

## À propos du bruit systémique

« Nous nous fions aux jugements professionnels, et nous faisons l'hypothèse implicite qu'un autre professionnel qualifié produirait un jugement à peu près identique. »

« Le bruit systémique est cinq fois plus important que ce que nous croyons, ou que le niveau que nous considérons comme tolérable. Sans un audit de bruit, nous ne l'aurions jamais compris. »

« Nous vivions dans l'illusion que nous partagions les mêmes jugements, une illusion d'accord. L'audit de bruit a fait voler en éclats cette illusion. »

« Le bruit systémique est un problème sérieux : il nous coûte des centaines de millions. »

« Dès lors qu'il y a jugement, il y a bruit – et plus qu'on ne l'imagine. »

CHAPITRE 3

# Les décisions singulières

Les études de cas que nous avons examinées portent sur des jugements qui sont rendus de façon répétée. Quelle est la juste peine dans une affaire de vol ? Quelle est la juste prime pour tel ou tel risque ? Si chaque cas est, à certains égards, unique, il s'inscrit dans un cadre, celui de *décisions récurrentes*. Les médecins diagnostiquent des patients, les juges entendent des prévenus, les comptables préparent des déclarations fiscales, les universités évaluent des dossiers de candidature : autant d'exemples de décisions récurrentes.

S'agissant des décisions récurrentes, le bruit peut être démontré par un audit de bruit, comme ceux que nous avons présentés dans les précédents chapitres. Quand des individus considérés comme interchangeables prennent chaque jour des décisions dans des cas similaires, il est facile de définir et de mesurer la variabilité indésirable. Il semble bien plus difficile, voire même impossible, d'appliquer la notion de bruit à une autre catégorie de jugements, celle des *décisions singulières*.

Considérons, par exemple, la crise mondiale due à la soudaine résurgence du virus Ebola en 2014. En Afrique de l'Ouest, les patients infectés mouraient en grand nombre. Compte tenu des flux de voyageurs, les projections indiquaient que l'épidémie se propagerait rapidement sur

tous les continents et toucherait l'Amérique du Nord et l'Europe de façon particulièrement grave. Aux États-Unis, on entendait des appels insistants pour interdire les vols en provenance des zones touchées, voire pour fermer les frontières. La pression politique en ce sens était forte, et des personnes influentes et bien informées y étaient favorables.

Le président Barack Obama était confronté à l'une des décisions les plus difficiles de sa présidence – une décision qu'il n'avait jamais rencontrée et qu'il ne rencontrerait plus. Il refusa de fermer les frontières. Il décida, en revanche, d'envoyer trois mille personnes, militaires et personnels de santé, en Afrique de l'Ouest. Il s'attacha par ailleurs à réunir une coalition internationale de pays qui, malgré leurs divergences, utilisèrent leurs ressources et leur expertise pour s'attaquer à la racine du problème.

## *Récurrence et singularité*

Les décisions qui ne sont prises qu'une seule fois, comme celle du président Obama sur le dossier Ebola, sont *singulières* : elles ne sont pas prises de manière récurrente par le même individu ou la même équipe ; il n'existe pas pour elles de solution toute faite ; et elles ont des caractéristiques proprement uniques. Face à Ebola, le président Obama et son équipe ne pouvaient pas vraiment s'appuyer sur des précédents. Les décisions politiques cruciales ou les choix les plus décisifs des états-majors militaires sont de bons exemples de décisions singulières.

Dans le domaine privé, les décisions que nous prenons quand nous choisissons un emploi, quand nous achetons une maison ou quand nous nous marions sont singulières. Même si ce n'est pas notre premier emploi, notre premier logement ou notre premier mariage, et même si d'innombrables personnes ont été avant nous confrontées à ces choix,

chaque décision nous paraît unique. De même, dans le monde économique, les dirigeants d'entreprise doivent souvent prendre des décisions qui leur semblent singulières : faut-il lancer une innovation susceptible de bouleverser notre marché ? fermer les bureaux pendant une épidémie ? ouvrir une filiale dans un nouveau pays ? céder à un gouvernement étranger qui veut réglementer notre activité ?

Certes, il n'y a pas entre les décisions singulières et les décisions récurrentes une différence de nature, mais plutôt un continuum. Il arrive aux assureurs de traiter d'affaires qui leur semblent sortir vraiment de l'ordinaire. Inversement, si vous achetez une maison pour la quatrième fois de votre vie, vous avez probablement commencé à considérer l'achat immobilier comme une décision récurrente. Mais les exemples extrêmes montrent clairement que la différence est significative. Déclarer la guerre est une chose, procéder à la revue budgétaire annuelle en est une autre.

## *Le bruit dans les décisions singulières*

Les décisions singulières sont traditionnellement considérées comme un objet d'étude bien distinct des jugements récurrents que produisent quotidiennement des employés interchangeables. Si les sciences humaines et sociales se sont souvent intéressées aux décisions récurrentes, l'étude des décisions singulières à forts enjeux a longtemps été la chasse gardée des historiens et des gourous du management.

Les approches des deux domaines sont bien différentes. L'analyse des décisions récurrentes a souvent une dimension statistique : le psychologue ou le sociologue étudient un ensemble de décisions similaires, afin d'y discerner des schémas communs, d'en identifier les régularités et d'en évaluer la qualité. En revanche, l'analyse des décisions singulières adopte le plus souvent une approche causale : elle se fait *a posteriori* et s'attache

surtout à identifier l'enchaînement des causes et des effets pour expliquer ce qui s'est passé. Les analyses historiques, comme les études de cas portant sur les succès ou les échecs des entreprises, s'attachent généralement à expliquer comment un jugement fondamentalement unique a été produit.

S'agissant des décisions singulières, l'étude du bruit soulève une difficulté particulière. Nous avons défini le bruit comme la variabilité indésirable des jugements relatifs au même problème. Or, puisque, par définition, les problèmes singuliers ne se répètent jamais exactement, cette définition ne leur est pas applicable. L'histoire ne repasse jamais les plats : nous ne pourrons jamais comparer la décision d'Obama dans la crise Ebola de 2014 avec les réactions d'autres présidents américains au même problème et au même moment (même si l'on peut, bien sûr, toujours spéculer). Vous pouvez accepter que nous comparions votre décision d'épouser l'être aimé à celle prise par d'autres amoureux avant vous, mais cette comparaison n'aura sans doute à vos yeux guère de sens, et en tout cas rien à voir avec celle que nous avons faite entre deux estimations d'une même prime d'assurance : vous et votre conjoint êtes bien sûr uniques. Il n'existe aucun moyen direct de démontrer la présence de bruit dans une décision singulière.

Pourtant, les mêmes facteurs qui produisent du bruit dans les décisions récurrentes sont bien présents, aussi, dans les décisions singulières. Souvenez-vous du stand de tir : si l'équipe C était sujette au bruit, c'est peut-être que les tireurs avaient réglé leur fusil de différentes manières. Ou peut-être, simplement, avaient-ils tous la main tremblante. Quoi qu'il en soit, si nous n'avions observé que le premier tireur de l'équipe, nous n'aurions eu aucune idée du bruit qui affectait son équipe. Pourtant, les mêmes sources de bruit n'en auraient pas moins été présentes. De même, quand vous prenez une décision une seule fois, il vous faut imaginer qu'une autre personne, à votre place, n'arriverait pas à la même conclusion, quand bien même elle serait aussi compétente que vous et partagerait vos objectifs

et vos valeurs. Et même si vous êtes seul à décider, il vous faut reconnaître que des changements apparemment insignifiants dans les détails de la situation ou dans le processus de prise de décision auraient tout à fait pu vous conduire à une décision différente.

Autrement dit, nous ne pouvons pas mesurer le bruit dans une décision singulière, mais si nous raisonnons de façon contrefactuelle, nous savons avec certitude qu'il y en a. De même que la main tremblante du tireur implique qu'un tir unique *aurait pu* toucher un autre point de la cible, le bruit qui affecte les décideurs et le processus de décision implique que la décision singulière *aurait pu* être différente.

Pour illustrer ce raisonnement contrefactuel, considérons tous les facteurs qui affectent une décision singulière. Si les experts chargés d'analyser la menace Ebola et de préparer des plans pour y répondre avaient été des personnes différentes, avec des parcours de vie et des expériences professionnelles différents, la proposition qu'ils ont soumise au président Obama aurait-elle été identique ? Si les mêmes faits avaient été présentés de façon légèrement différente, la conversation se serait-elle déroulée de la même façon ? Si les acteurs clés avaient été d'une autre humeur, ou si la réunion décisive avait eu lieu pendant une tempête de neige, la décision finale aurait-elle été changée ? Vue sous cet angle, la décision singulière ne semble plus aussi déterminée. Il devient possible de penser qu'en fonction des variations d'un certain nombre de facteurs, pour certains si insignifiants que nous n'en sommes même pas conscients, la décision aurait pu être très différente.

Un autre exemple de raisonnement contrefactuel nous est fourni par la crise du Covid-19. Même quand le virus les a frappés à peu près en même temps et de la même façon, les différents pays touchés ont réagi de manière extrêmement contrastée. Cette variation montre clairement la présence de bruit dans les processus de décision de ces pays. Mais imaginons ce qui se serait passé si le virus n'était apparu que dans un seul pays. À l'évidence,

nous n’aurions pas pu observer de variabilité. Pourtant, les décisions des dirigeants de ce pays hypothétique auraient été affectées par les mêmes sources de bruit que celles des dirigeants qui ont géré la crise dans le monde réel.

En somme, quand une décision est singulière, nous ne pouvons évidemment pas en observer la variabilité. Mais ce n’est pas parce que nous sommes incapables d’observer de la variabilité que la décision est exempte de bruit.

## *Réduire le bruit dans les décisions singulières*

Cette manière d’appréhender le bruit dans les décisions singulières peut sembler quelque peu théorique, mais elle est essentielle. La raison en est simple : si le bruit affecte les décisions singulières de la même manière que les décisions récurrentes, alors les stratégies qui réduisent le bruit dans celles-ci devraient aussi améliorer la qualité de celles-là.

Cette observation est beaucoup plus contre-intuitive qu’il n’y paraît. Quand vous devez prendre une décision qui vous semble unique en son genre, votre instinct est probablement de la traiter comme telle, en faisant abstraction de toute comparaison avec d’autres situations. Certains théoriciens soutiennent même que les règles de la pensée probabiliste sont totalement inadaptées aux décisions singulières prises dans une situation d’incertitude et que celles-ci nécessitent une approche entièrement différente.

Le raisonnement que nous tenons ici suggère le contraire. Du point de vue de la réduction du bruit, *une décision singulière est une décision récurrente qui n’est prise qu’une seule fois*.

Il y a des différences réelles entre décisions récurrentes et décisions singulières. Et nous sommes tentés de les aborder d'une manière radicalement différente. Pourtant, qu'un problème se présente cent fois ou une seule, votre objectif devrait être le même : réduire à la fois le biais et le bruit qui affectent votre décision. Et les mêmes pratiques qui réduisent l'erreur dans les décisions récurrentes devraient pouvoir vous aider à améliorer celles que vous ne prenez qu'une seule fois.

### À propos des décisions singulières

« Cette décision est unique et nous ne pourrons jamais en mesurer le bruit ; mais elle est quand même affectée par de nombreux facteurs qui ne devraient pas l'influencer. »

« Il est aussi important de combattre le bruit dans une décision singulière que dans une décision récurrente : une décision singulière est une décision récurrente que nous ne prenons qu'une seule fois. »

PARTIE II

# Votre esprit est un instrument de mesure

En sciences comme dans le langage courant, *mesurer*, c'est utiliser un instrument pour attribuer à un objet une valeur choisie sur une échelle. La longueur d'un tapis, par exemple, se mesure en centimètres avec un mètre mesureur ; la température se mesure en degrés Fahrenheit ou Celsius à l'aide d'un thermomètre.

Produire un jugement est un acte similaire. Quand un juge décide du nombre de mois de réclusion qui lui semble approprié pour sanctionner un crime, il choisit une valeur sur l'échelle des peines de prison. C'est aussi ce que fait le souscripteur de la compagnie d'assurances quand il fixe une valeur en dollars pour assurer un risque, ou le médecin quand il fait un diagnostic. L'échelle n'est pas toujours chiffrée : quand on déclare un accusé « coupable au-delà de tout doute raisonnable », quand on qualifie une lésion de « mélanome avancé », ou quand on suggère qu'« une opération chirurgicale est fortement recommandée », ces affirmations sont aussi des jugements.

Le jugement peut donc être considéré comme *une mesure dont l'instrument est un esprit humain*. Parler de mesure suggère implicitement

que le but poursuivi est la justesse : une mesure doit s'approcher le plus possible de la vérité, c'est-à-dire minimiser l'erreur. De même, l'objectif d'un jugement est d'être juste. Ce n'est ni d'impressionner, ni de prendre position, ni de persuader. Il est important de souligner que le concept de jugement tel que nous l'utilisons ici est emprunté à la littérature psychologique, et que cette définition est beaucoup plus étroite que celle du langage quotidien. *Juger*, au sens où nous l'entendons ici, n'est pas synonyme de *penser* ; et *produire des jugements justes ou corrects* n'est pas la même chose qu'*avoir un bon jugement*.

Tel que nous le définissons, le jugement – comme la mesure – est une conclusion. Il peut être résumé d'un mot ou d'une formule. Si un analyste du renseignement écrit un long rapport dont la conclusion est que tel régime politique est instable, c'est cette conclusion, et non l'ensemble du rapport, qui constitue un jugement. Ajoutons que le jugement (là encore, comme la mesure) renvoie à la fois à une opération et à son produit : le mot désigne à la fois l'activité mentale qui consiste à juger et la conclusion elle-même. Enfin, nous utiliserons ici le mot *juge* dans le sens technique que la psychologie du jugement donne à ce terme : le « juge » est l'individu qui produit des jugements, même quand il n'a aucun rapport avec le système judiciaire. (Les définitions des termes techniques comme *jugement*, *erreur*, *bruit* et *biais* sont reprises dans le glossaire de cet ouvrage.)

L'objectif de tout jugement, comme de toute mesure, est donc la justesse. Mais la perfection n'est pas de ce monde, et l'exactitude parfaite demeure éternellement inaccessible. Dans la mesure scientifique, et *a fortiori* dans le jugement, il reste toujours de l'erreur. Une partie de cette erreur est du biais ; une autre, du bruit.

Pour comprendre les contributions respectives du bruit et du biais à l'erreur, nous vous invitons à jouer à un jeu qui vous prendra moins d'une minute. Si votre smartphone est équipé d'un chronomètre, il a sans doute une fonction *lap* ou « temps par tour » qui permet de mesurer des

intervalles successifs de temps sans arrêter le chronomètre ni même le consulter. Le but du jeu que nous vous proposons est de mesurer cinq temps successifs d'exactement dix secondes sans regarder votre smartphone. Si vous le souhaitez, vous pouvez observer un intervalle de dix secondes plusieurs fois avant de commencer. Prêts ? Partez !

Regardez maintenant les différents intervalles de temps enregistrés sur votre smartphone. (L'appareil lui-même n'est pas exempt de bruit, mais on peut le considérer comme négligeable.) Vous constaterez qu'ils ne sont pas exactement de dix secondes et qu'ils varient sans doute substantiellement. Vous avez essayé de reproduire exactement le même intervalle de temps, mais sans succès. La variabilité que vous n'avez pas pu contrôler est un exemple de bruit.

Cette observation n'est guère surprenante, car le bruit est omniprésent en physiologie et en psychologie. La variabilité entre individus est un fait biologique de base : deux pois dans une cosse ne sont jamais totalement identiques. On retrouve aussi toutes sortes de variabilités chez l'individu. Votre rythme cardiaque n'est pas tout à fait régulier. Vous ne pouvez pas répéter le même geste avec une précision parfaite. Et si vous faites un bilan auditif, certains sons seront si bas que vous ne les entendrez pas, et d'autres si forts que vous les entendrez chaque fois ; mais il y en aura aussi que vous entendrez certaines fois et pas d'autres.

Regardez maintenant les cinq temps enregistrés sur votre smartphone. Voyez-vous apparaître une tendance ? Les cinq temps sont-ils tous inférieurs à dix secondes, par exemple, ce qui pourrait signifier que votre « horloge » interne avance ? Ou sont-ils tous supérieurs à dix secondes ? Dans cette tâche simple, le biais est la différence, positive ou négative, entre la *moyenne* de vos temps chronométrés et la cible de dix secondes. Le bruit est la *variabilité* de vos mesures, analogue à la dispersion des tirs sur la cible. En statistique, la mesure la plus courante de la variabilité est l'*écart-*

*type*, et c'est de celui-ci que nous nous servirons pour mesurer le bruit dans les jugements[1].

On peut considérer la plupart des jugements, et en particulier les *jugements prédictifs*, comme analogues à la mesure que vous venez de prendre. Quand nous faisons une prédiction, nous essayons d'approcher d'une valeur vraie ou juste. Le prévisionniste en économie souhaite approcher autant que possible de la véritable valeur de la croissance du produit intérieur brut de l'année à venir. Le médecin souhaite faire le bon diagnostic. (Notez que la « prédiction », au sens technique où nous utilisons le mot dans ce livre, n'implique pas de prédire l'avenir : le diagnostic d'un état médical présent est une « prédiction ».)

L'analogie entre le jugement et la mesure joue un rôle essentiel dans ce livre, car elle nous aidera à expliquer le rôle du bruit dans l'erreur. Les personnes qui produisent des jugements prédictifs sont comparables au tireur qui vise le centre de la cible ou au physicien qui mesure la masse d'une particule. La présence de bruit dans leurs jugements implique de l'erreur : dès lors que le but du jugement est de s'approcher d'une valeur vraie, deux jugements différents ne peuvent pas être tous les deux exacts. Bien sûr, comme les instruments de mesure, les individus peuvent être plus ou moins précis. Certains, moins compétents ou mal formés, font plus d'erreurs que d'autres. Mais, pas plus que les instruments de mesure, les individus ne sont jamais parfaits. Leurs jugements sont toujours entachés d'erreurs, qu'il nous faut comprendre et mesurer.

Il va sans dire que la plupart des jugements professionnels sont infiniment plus complexes que la mesure d'un intervalle de temps. Dans le chapitre 4, nous définirons les différents types de jugements professionnels et verrons quelle est leur finalité. Les chapitres 5 et 6 présenteront la manière de mesurer l'erreur, et d'en analyser les deux composantes : le biais, sa moyenne ; et le bruit, sa dispersion. Enfin, les chapitres 7 et 8 commenceront à présenter les circonstances qui donnent naissance au bruit.

Le message qui se dégage de ces chapitres est simple : le jugement humain, comme un instrument de mesure, est imparfait. Il est sujet à la fois au biais et au bruit. Pourquoi et dans quelles proportions ? C’est ce que nous allons voir.

CHAPITRE 4

# Questions de jugement

Ce livre traite des jugements professionnels, au sens large du terme. Nous partons du principe que le professionnel qui produit un jugement est raisonnablement compétent, et que son but est d'« avoir bon » ou de « trouver juste ». Mais le concept même de jugement implique de reconnaître, fût-ce à contrecœur, qu'on ne peut jamais être certain qu'un jugement soit parfaitement exact.

Songeons à des expressions comme « c'est une question de jugement », « ça se discute », ou « c'est un jugement subjectif ». Quand on affirme que « le soleil se lèvera demain » ou que « la formule du chlorure de sodium est NaCl », ce ne sont pas des « questions de jugement ». Des gens raisonnables doivent être parfaitement d'accord sur ces deux points. Ce qui caractérise au contraire une question de jugement, c'est une certaine incertitude, et cette incertitude implique que des gens raisonnables et compétents puissent y apporter des réponses différentes.

Toutefois, il y a une limite au niveau de désaccord acceptable. Le mot *jugement* est d'ailleurs utilisé principalement là où les gens pensent qu'ils *devraient* être d'accord. Les questions de jugement ne sont pas des questions d'opinion ou de goût, pour lesquelles des différences persistantes sont normales et acceptées. Les dirigeants de la compagnie d'assurances,

qui trouvaient intolérable le bruit dans l'évaluation des sinistres, n'auraient rien trouvé à redire si leurs experts s'étaient déchirés sur les mérites relatifs des Beatles et des Rolling Stones, ou sur ceux du saumon et du thon.

En d'autres termes, les questions de jugement, et en particulier de jugement professionnel, occupent une position intermédiaire entre les questions de fait ou de calcul, d'un côté, et les questions de goût et d'opinion, de l'autre. Elles sont caractérisées par une *anticipation de désaccord limité*.

Définir la limite de ce désaccord, c'est-à-dire ce qui sera considéré comme acceptable, est en soi un jugement subjectif. À l'évidence, tout dépend de la difficulté du problème. Dans les concours de vins, les juges s'entendent rarement pour distribuer des médailles, mais il est fréquent que les vins les plus médiocres leur inspirent un dédain unanime[1]. Et même sur des problèmes difficiles, le consensus se fera facilement sur des jugements manifestement absurdes. Des magistrats qui jugent une affaire ordinaire de vol aggravé ne seront sans doute pas précisément d'accord sur la sentence, mais ils conviendront tous qu'une amende d'un dollar et la prison à perpétuité sont deux peines déraisonnables.

## *L'expérience du jugement : un exemple*

Avant de poursuivre, nous vous proposons de faire une petite expérience de jugement, qui vous permettra, si vous vous livrez à l'exercice jusqu'au bout, de tirer davantage profit du reste du chapitre.

> *Vous faites partie de l'équipe chargée d'évaluer des candidats pour le poste de président-directeur général d'une entreprise financière régionale qui a des résultats moyens et qui doit faire face à une concurrence accrue. Il vous est demandé d'évaluer la probabilité qu'un*

*candidat nommé Michael Gambardi réussisse dans ce poste, à un horizon de deux ans. « Réussir » signifie simplement ici que le candidat sera toujours P-DG à l'issue de ces deux années. Vous noterez la probabilité sur une échelle de 0 (impossible) à 100 (certain).*

> *Michael Gambardi a 37 ans. Diplômé de la Harvard Business School il y a douze ans, il a occupé depuis plusieurs postes. Très tôt, il a créé et investi dans deux start-up qui, faute de lever des fonds suffisants, ont périclité. Il est entré ensuite dans une grande compagnie d'assurances et s'est hissé rapidement au poste de directeur régional des opérations pour l'Europe. À ce titre, il a lancé et géré une réorganisation qui a permis de réduire de façon importante le délai de traitement des sinistres. Ses collègues et ses subordonnés le décrivent comme efficace, mais aussi dominateur et rugueux ; il y a eu un turnover important dans l'équipe de direction quand il était en fonctions. Depuis deux ans, il est P-DG d'une firme financière de taille moyenne qui, à son arrivée, était au bord de la faillite. Ses collègues le créditent d'avoir stabilisé l'entreprise, mais notent aussi qu'il est difficile de travailler avec lui. Il a exprimé le désir de s'attaquer à un nouveau défi. Les spécialistes en ressources humaines qui se sont entretenus avec lui il y a quelques années lui ont donné de très bonnes notes pour l'énergie et la créativité, mais ont aussi jugé qu'il pouvait être arrogant et quelquefois tyrannique.*

*Rappelons que Michael Gambardi est candidat à un poste de P-DG d'une entreprise financière régionale qui réussit moyennement et qui doit faire face à une concurrence accrue. Quelle est la probabilité, s'il est engagé, qu'il soit encore à son poste au bout de deux ans ? Choisissez un chiffre entre 0 et 100 avant de poursuivre la lecture. Relisez la description si besoin est.*

Si vous avez pris l'exercice au sérieux, vous l'avez sans doute trouvé difficile. Les informations fournies sont nombreuses et contradictoires. Il vous a fallu faire un effort pour vous forger une impression cohérente, sans laquelle vous ne pouvez pas produire un jugement. Pour construire cette impression, vous avez dû vous concentrer sur les détails qui vous paraissaient les plus importants, et vous en avez sans doute ignoré d'autres. Si l'on vous demandait d'expliquer le chiffre que vous avez choisi, vous évoqueriez sans doute quelques faits qui vous semblent sortir du lot, mais vous seriez bien en peine de justifier pleinement votre jugement.

Le processus mental dont vous venez de faire l'expérience illustre plusieurs caractéristiques importantes de l'opération que nous appelons « jugement ».

- D'abord, parmi tous les indices que contient la description de Michael (et qui ne révèlent, bien sûr, qu'une petite partie de ce que vous auriez idéalement besoin de savoir), vous avez prêté attention à certains éléments plutôt que d'autres, sans être toujours pleinement conscient des choix que vous faisiez. Avez-vous remarqué que Gambardi est un nom italien ? Vous rappelez-vous les études qu'il a faites ? Dans cet exercice, comme généralement dans la vie, les informations disponibles sont trop nombreuses pour que vous puissiez facilement les conserver toutes en mémoire. Selon toute vraisemblance, ce que vous avez retenu n'est pas identique à ce qu'auront retenu d'autres lecteurs. Ainsi, l'attention sélective et la mémoire sélective sont une première source de variabilité entre les personnes.

- Vous avez ensuite intégré informellement ces indices pour vous former une impression générale sur les perspectives de Gambardi. Le mot essentiel ici est l'adverbe *informellement*. Vous n'avez pas fait de plan pour répondre à la question. Sans en être pleinement conscient, votre esprit a œuvré à construire une impression cohérente des forces et des

faiblesses du candidat et des difficultés qui l'attendaient. Ce caractère informel vous a permis de travailler rapidement. Mais il a aussi produit de la variabilité : tandis qu'un processus formalisé garantit des résultats identiques (comme quand on additionne une colonne de chiffres), dans une opération informelle, le bruit est inévitable.

- Vous avez enfin converti cette impression générale en un chiffre sur l'échelle des probabilités de réussite. Votre aptitude à faire correspondre un chiffre entre 0 et 100 à une impression est un processus étonnant (nous y reviendrons en détail dans le chapitre 14). Là encore, vous ne savez pas vraiment pourquoi vous avez répondu comme vous l'avez fait : si vous avez choisi 65, pourquoi pas 61 ou 69 ? Vraisemblablement, un chiffre vous est venu à l'esprit à un certain moment. Vous vous êtes demandé si ce chiffre vous semblait acceptable ; dans le cas contraire, un autre chiffre vous est venu à l'esprit. Cette partie du processus de jugement est, elle aussi, source de variabilité entre les personnes.

Répétons-le : chacune des étapes de ce processus de jugement complexe génère de la variabilité. On peut donc s'attendre à ce que les réponses concernant la probabilité de succès de Michael Gambardi soient très variables. Et c'est bien ce que vous constaterez si vous demandez à des amis de faire l'exercice : leurs estimations seront très dispersées. Nous avons par exemple soumis le cas Gambardi à 115 étudiants de MBA, et leurs réponses allaient de 10 à 95 %. Cela fait beaucoup de bruit !

Notons au passage que l'exercice du chronomètre et le cas Gambardi illustrent deux types distincts de bruit. La variabilité des jugements sur des essais successifs est un exemple de bruit propre à un seul juge (vous) ; la variabilité des jugements sur Michael Gambardi est un exemple de bruit entre plusieurs juges. En termes techniques, le premier problème illustre la variabilité *intra-individuelle*, et le second la variabilité *interindividuelle*.

# *Le signal intérieur*

Votre réponse au problème Gambardi est un exemple de ce que nous avons appelé un jugement prédictif. Mais il présente une différence importante avec d'autres jugements prédictifs que nous avons cités. Si vous n'êtes pas d'accord avec un ami sur la température qu'il fera demain à Bangkok, sur le score du match de football de ce soir, ou sur le résultat de la prochaine élection présidentielle, le moment viendra où vous saurez qui avait raison. Mais si vous n'êtes pas d'accord sur le cas Gambardi, le temps *ne vous dira pas* qui avait vu juste.

À cela, il existe une première raison simple : Gambardi n'existe pas. Mais même si la question concernait une personne réelle, et même si nous connaissions son sort au bout de deux ans, cette information ne vous dirait pas quelle était la probabilité de succès de Gambardi *au moment du jugement*. Un jugement de probabilité ne peut être ni confirmé ni infirmé par les faits (sauf s'il était de 0 % ou de 100 %). Si, par exemple, un événement dont vous estimiez qu'il avait 90 % de chances de se produire ne se produit pas, cela ne signifie pas nécessairement que votre jugement de probabilité était mauvais. Car ce qu'on évalue à raison comme n'ayant que 10 % de chances d'advenir doit tout de même se produire de temps en temps – précisément, même, dans 10 % des cas ! L'exercice Gambardi est donc un exemple de jugement prédictif *invérifiable*, et ce pour deux raisons distinctes : Gambardi est un personnage fictif ; et la réponse est de nature probabiliste.

Un grand nombre de jugements professionnels sont invérifiables. Sauf erreur flagrante, les assureurs ne sauront jamais, par exemple, si le prix d'une police d'assurance donnée est surévalué ou sous-évalué. D'autres prévisions ne sont pas vérifiables parce qu'elles sont conditionnelles : « si nous déclarons la guerre, nous subirons une défaite écrasante » est une prédiction importante, mais son contenu même rend probable (et fort

souhaitable) qu'elle ne soit jamais vérifiée. D'autres prévisions sont à trop long terme pour que les professionnels qui les ont faites aient à en rendre compte un jour : songez, par exemple, aux estimations de l'élévation des températures d'ici à la fin du XXIe siècle.

Pourtant, dans l'exercice Gambardi, le fait que la prédiction soit invérifiable a-t-il changé la manière dont vous l'avez approchée ? Vous êtes-vous demandé, par exemple, si Gambardi était réel ou fictif ? Ou si le résultat vous serait donné plus tard dans le texte ? Avez-vous réfléchi sur le fait que, même si c'était le cas, cette révélation ne vous donnerait pas la réponse à la question qui vous était posée ? Probablement pas, parce que ces considérations ne vous ont pas semblé pertinentes au moment de répondre à la question.

La vérifiabilité ne change pas l'expérience du jugement. Certes, on peut réfléchir davantage à un problème dont on sait que la solution sera bientôt connue, ne serait-ce que par crainte de faire une erreur grossière et embarrassante. Inversement, on peut aussi refuser de consacrer des efforts à un problème tellement hypothétique qu'il en devient absurde (« si Gambardi avait trois jambes et pouvait voler, ferait-il un meilleur P-DG ? »). Mais, de façon générale, notre approche des problèmes hypothétiques, pour autant qu'ils soient plausibles, est sensiblement la même que celle que nous appliquons aux problèmes réels. Cette similitude n'est pas sans importance pour la recherche en psychologie, qui soumet souvent à ses sujets d'expérience des problèmes inventés de toutes pièces.

Cette observation soulève un problème. Puisqu'il est invérifiable, le jugement qui vous était demandé dans l'exercice Gambardi n'a pas à proprement parler de « résultat ». D'ailleurs, nous l'avons vu, vous ne vous êtes sans doute pas demandé s'il en aurait jamais un. Votre objectif en formulant ce jugement ne pouvait donc pas être de trouver ce résultat, ni de vous en approcher au plus près. En réalité, votre effort pour « juger juste » a consisté à chercher un chiffre en lequel vous aviez suffisamment confiance

pour en faire votre réponse. Bien sûr, cette confiance n'est pas totale : vous êtes moins sûr de votre estimation que vous ne l'êtes du fait que deux et deux font quatre. Vous avez conscience d'une certaine incertitude (et, nous y reviendrons, vous la sous-estimez sans doute). Mais, à un moment donné, vous avez conclu que vous aviez assez réfléchi et vous avez donné votre réponse.

Qu'est-ce qui vous a donné le sentiment que votre jugement était juste, ou du moins assez juste pour en faire votre réponse ? Notre thèse est que ce sentiment est un *signal intérieur*, qui commande la fin de l'opération de jugement, et qui ne dépend d'aucune validation externe. Le moment où votre réponse vous a paru juste est celui où elle vous a semblé suffisamment cohérente avec les données du problème. Une réponse de 0 % ou 100 % n'aurait pas pu vous donner ce sentiment : des jugements aussi extrêmes ne sont pas cohérents avec les données ambiguës, brouillonnes et contradictoires qui vous étaient fournies. Mais le chiffre sur lequel vous vous êtes arrêté, quel qu'il soit, vous a apporté le sentiment de cohérence dont vous aviez besoin. Le but véritable du jugement, tel que vous en avez fait l'expérience, est de parvenir à ce sentiment de cohérence.

Le signal intérieur et le sentiment de cohérence qui le déclenche font partie intégrante de l'expérience du jugement. Ils ne dépendent pas d'un résultat réel. C'est ce qui permet au signal intérieur de résonner de la même manière quand les jugements portent sur des événements réels et vérifiables et quand ils impliquent des personnages fictifs comme Michael Gambardi. Et c'est ce qui rend notre expérience subjective du jugement pratiquement identique dans les deux cas.

## *Le résultat et le processus*

Le caractère vérifiable ou non d'un jugement ne change pas l'expérience du jugement au moment où celui-ci est produit. Mais il peut, bien sûr, changer l'*évaluation* que l'on en fait *a posteriori*.

Quand un jugement est vérifiable, un observateur objectif peut l'évaluer à l'aide d'une mesure simple de l'erreur : la différence entre le jugement et le résultat. Si un météorologiste avait annoncé pour aujourd'hui une température de 21 °C alors qu'elle est en réalité de 18 °C, il a fait une erreur de + 3 °C. Toutefois, ce raisonnement n'est évidemment pas applicable aux jugements invérifiables (comme celui du cas Gambardi), puisqu'ils n'ont pas à proprement parler de résultat. Dès lors, comment peut-on parler de bon ou de mauvais jugement ?

Il faut, pour cela, faire appel à une autre manière d'évaluer les jugements, et qui s'applique aussi bien aux jugements invérifiables qu'aux jugements vérifiables : évaluer le *processus* de jugement. Quand nous parlons de bon ou de mauvais jugement, nous pouvons parler soit du résultat (par exemple, le chiffre que vous avez produit dans le cas Gambardi), soit du processus (ce que vous avez fait pour arriver à ce chiffre).

Une méthode courante pour évaluer un processus de jugement consiste à observer les résultats qu'on obtient lorsqu'on applique ce processus à un grand nombre de cas. Prenons l'exemple d'un politologue qui a attribué des probabilités de victoire à un grand nombre de candidats à des élections locales. Il estime que cent de ces candidats ont 70 % de chances de gagner. Si soixante-dix d'entre eux l'emportent effectivement, nous pourrons conclure que la méthode du politologue pour calibrer ses jugements de probabilité est bonne. Pris dans leur ensemble, ses jugements deviennent vérifiables et peuvent ainsi être évalués, même si aucun jugement de probabilité, pris isolément, ne peut être déclaré bon ou mauvais. De la même manière, pour déceler, dans un processus de sélection, la présence

d’un biais favorable ou hostile à telle ou telle catégorie de candidats, on devra examiner des statistiques portant sur un nombre important de cas.

Pour évaluer un processus de jugement, on peut aussi vérifier qu’il se conforme aux principes de la logique et à la théorie des probabilités. Un vaste corpus de recherche sur les biais cognitifs du jugement suit cette méthode.

S’intéresser au processus du jugement plutôt qu’à son résultat permet d’évaluer la qualité de jugements invérifiables, notamment les jugements portant sur des problèmes fictifs ou les prévisions à long terme. Même si nous ne pouvons pas les comparer à un résultat connu, nous pouvons dire s’ils ont été ou non produits de manière correcte. Et, quand nous irons au-delà de l’évaluation des jugements pour aborder la question de leur *amélioration*, c’est aussi au processus que nous nous intéresserons. Toutes les procédures que nous recommanderons dans ce livre pour réduire le biais et le bruit tendent vers le même but : adopter, dans chaque situation de jugement, le processus qui produirait la plus petite erreur possible si on l’appliquait à un ensemble de jugements similaires.

Nous venons de distinguer deux manières d’évaluer le jugement : en le comparant à un *résultat* et en évaluant la qualité du *processus* qui y a abouti. Il est important de noter que, quand on s’intéresse à un cas unique et quand le résultat de celui-ci est vérifiable, ces deux manières d’évaluer le jugement peuvent conduire à des conclusions opposées. Un prévisionniste compétent et attentif, utilisant les meilleurs outils et techniques possibles, se trompera souvent dans sa prévision du taux d’inflation trimestriel : le meilleur processus possible ne garantit pas le bon résultat. Inversement, la pire des méthodes n’est jamais à l’abri d’un succès : un chimpanzé qui lance des fléchettes tombera quelquefois sur le chiffre juste.

Face à ce dilemme, les théoriciens de la décision ont fait leur choix : ce qui compte à leurs yeux, ce n’est jamais le résultat dans un cas unique, mais bien la qualité du processus. Intellectuellement, cette position est

séduisante ; c'est même sans doute la seule tenable. En pratique, toutefois, il faut convenir que la vie réelle n'en tient à peu près aucun compte. Dès lors que des jugements sont vérifiables, ceux qui font profession de les produire sont évalués sur leur capacité à s'approcher au plus près des résultats. D'ailleurs, si vous demandez à un prévisionniste quel est son objectif, il vous le dira sans hésiter : produire une prévision qui s'avérera juste.

Pour résumer, il y a au moins trois manières de définir ce vers quoi tend le jugement. Ce que les gens *disent* généralement rechercher dans un jugement vérifiable, et ce dont on les tient le plus souvent responsables, c'est une correspondance (ou une forte proximité) avec le *résultat*. Ce qu'ils essaient réellement de *ressentir*, indépendamment de la vérifiabilité, c'est un *signal intérieur*, qui conclut l'opération de jugement, et qui est suscité, en l'absence de tout résultat, par la cohérence entre les données du problème et le jugement produit. D'un point de vue normatif, enfin, ce qu'ils *devraient* essayer de trouver, c'est le meilleur *processus de jugement*, c'est-à-dire celui qui serait susceptible de minimiser les erreurs sur un ensemble de cas similaires.

## *Les jugements évaluatifs*

Depuis le début de ce chapitre, nous parlons de jugements prédictifs. Et beaucoup des jugements dont nous traiterons sont effectivement de ce type. Certains, néanmoins, sont très différents. Rappelez-vous le premier chapitre, où nous évoquions le juge Frankel et les peines prononcées par les juges fédéraux. À l'évidence, condamner un criminel n'a rien d'une prédiction. C'est ce que nous appellerons un *jugement évaluatif* : en l'espèce, il s'efforce de trouver une correspondance entre la peine et la gravité du crime. Les œnologues qui distinguent les meilleurs vins ou les

critiques gastronomiques qui classent des restaurants produisent des jugements évaluatifs. Il en va de même des professeurs qui notent des dissertations, des juges dans les compétitions de patinage artistique, ou encore des commissions qui accordent des financements à des projets de recherche.

Il existe en particulier un type important de jugement évaluatif : les décisions qui nécessitent l'examen de multiples options présentant des avantages et des inconvénients entre lesquels il faut arbitrer. Prenons les recruteurs qui choisissent parmi des candidats à l'embauche, les équipes de direction qui décident de choix stratégiques, ou même les chefs d'État qui choisissent comment faire face à une épidémie. Bien sûr, chacune de ces décisions est alimentée par des jugements prédictifs. Pour décider, il faut d'abord prévoir comment chaque candidat se comportera durant sa première année d'embauche, quelle sera la réaction du marché à chacune des options stratégiques, ou à quelle vitesse l'épidémie se propagera si l'on ne fait rien pour la freiner. Mais la décision finale passe nécessairement par un arbitrage entre les avantages et les inconvénients des différentes options en présence, et cet arbitrage ne peut être résolu que par un jugement évaluatif [2].

Un jugement évaluatif reste un jugement : comme un jugement prédictif, il implique une anticipation de désaccord limité. Aucun magistrat digne de ce nom ne dirait, à propos de la peine qu'il vient de fixer : « C'est le chiffre que je préfère, et je me fiche que mes collègues pensent autrement. » Quant aux décideurs qui choisissent entre plusieurs options stratégiques, ils s'attendent naturellement à ce que des collègues qu'ils respectent, s'ils disposent des mêmes informations et partagent les mêmes objectifs, soient peu ou prou d'accord avec eux. Les jugements évaluatifs dépendent bien sûr pour partie des valeurs et des préférences des personnes qui les produisent, mais ce ne sont pas des affaires de goût ou d'opinion.

C'est précisément pour cette raison que la frontière entre le jugement prédictif et le jugement évaluatif est assez floue. Les individus qui

produisent des jugements n’ont souvent même pas conscience de cette distinction. Les juges qui fixent des peines ou les professeurs qui notent des dissertations réfléchissent longuement à leur tâche et s’efforcent de trouver la « bonne » réponse. Ils prennent confiance dans leurs jugements et dans les raisons qu’ils se donnent pour faire les choix qu’ils font. En cela, ils ne sont pas très différents des prévisionnistes ou des médecins qui formulent un diagnostic. Que l’on formule un jugement prédictif (« Quel succès aura ce nouveau produit ? ») ou un jugement évaluatif (« Mon assistant a-t-il été performant cette année ? »), on pense, on agit et on se justifie à peu près de la même façon.

## *Le bruit, toujours problématique*

Dans les jugements prédictifs, le bruit signale toujours que quelque chose ne va pas. Si deux médecins ne s’accordent pas sur un diagnostic, ou si deux chefs des ventes font des prévisions très différentes pour le trimestre prochain, il ne fait aucun doute qu’au moins l’un des deux jugements est dans l’erreur. La cause du désaccord peut, d’ailleurs, être un simple problème de compétence : si l’un des individus en question est insuffisamment qualifié, il se trompera souvent. Mais il y a bien d’autres sources de bruit. Et quelle qu’en soit la cause, le bruit peut être lourd de conséquences quand il affecte des jugements importants.

Le bruit dans les jugements évaluatifs est également problématique, mais pour d’autres raisons. Dans tout système dans lequel des juges considérés comme interchangeables se voient assigner des tâches de manière aléatoire, on s’attend à une certaine équité de traitement. L’existence de désaccords importants sur la même affaire va à l’encontre de cette attente. Si des juges différents, face au même prévenu, peuvent prononcer des peines très différentes, nous sommes dans le monde des

« cruautés arbitraires » que dénonçait le juge Frankel. Même des magistrats qui défendent l'individualisation des peines, et qui trouvent donc normal de ne pas s'accorder sur la peine précise à infliger dans chaque cas particulier, en conviendront aisément : des désaccords si grands qu'ils transforment le jugement en loterie sont problématiques. Il en va de même (avec des conséquences certes moins dramatiques) quand des notes très différentes sont données à la même dissertation, au même restaurant ou au même patineur – ou quand une personne souffrant de dépression obtient un congé maladie, tandis qu'une autre, souffrant du même mal, n'a que ses yeux pour pleurer.

Même quand l'équité n'est pas une préoccupation capitale, le bruit systémique reste problématique. Dans tout système qui produit des jugements évaluatifs, on s'attend à de la cohérence. Les jugements devraient refléter les valeurs et les priorités du système, et non les préférences, les humeurs ou les caprices de l'individu auquel on a affaire. Il y a quelque chose qui ne tourne pas rond dans un magasin où un client qui se plaint d'un ordinateur défectueux se fait intégralement rembourser tandis qu'un autre n'obtient que des excuses ; ou dans une entreprise où un employé obtient la promotion qu'il demande, tandis qu'un autre dont les performances sont identiques essuie un refus poli. Le bruit systémique est une incohérence ; et toute incohérence porte atteinte à la crédibilité du système.

## *Indésirable mais mesurable*

Pour mesurer le bruit, nous n'avons besoin que d'une chose : un ensemble de jugements différents du même problème. Nous n'avons pas besoin de connaître la valeur « juste » que ces jugements tentaient d'approcher. C'est ce qu'illustrait notre exemple introductif du stand de tir : quand on regarde

le dos de la cible, la dispersion des points d'impact reste parfaitement visible, même si la position du centre de la cible ne l'est pas. Dès lors que l'on sait que tous les tireurs visaient le centre de la cible, on peut aisément mesurer le bruit. C'est le principe de l'audit de bruit. Si nous demandons à tous nos prévisionnistes d'estimer les ventes du trimestre à venir, nous ne saurons rien – pour l'instant – du biais dans leurs prévisions, mais la dispersion de leurs chiffres nous donnera immédiatement une mesure du bruit.

---

Pour qui s'efforce d'améliorer la qualité des jugements, cette différence entre le biais et le bruit est capitale. Il peut sembler étrange de suggérer qu'on peut améliorer des jugements sans pouvoir vérifier s'ils sont ou non corrects. C'est pourtant tout à fait possible, si l'on commence par mesurer le bruit. Qu'il s'agisse d'un jugement prédictif, qui vise l'exactitude, ou d'un jugement évaluatif, qui implique un arbitrage plus complexe entre des valeurs différentes, le bruit est toujours indésirable et souvent mesurable. Et une fois qu'il est mesuré, comme on le verra dans la cinquième partie de ce livre, il est généralement possible de le réduire.

## À propos des jugements professionnels

« C’est une question de jugement. On ne peut pas s’attendre à ce que les gens soient tous parfaitement d’accord. »

« Oui, c’est une question de jugement, mais certains jugements sont si absurdes qu’ils sont forcément erronés. »

« Vous vous sentez confiant dans ce jugement ; mais cette confiance est-elle fondée sur des faits objectifs, ou seulement sur un signal intérieur ? »

« J’estimais les chances de ce candidat à 80 %, et il a perdu – mais cela ne veut pas dire que j’avais tort. Un jugement probabiliste n’est pas vérifiable. »

« Toute décision nécessite à la fois des jugements prédictifs et des jugements évaluatifs. D’ailleurs, qu’ils soient prédictifs ou évaluatifs, notre expérience du jugement n’est pas très différente. »

« Le bruit est toujours nuisible. Ce n’est pas parce que nous ne savons pas lequel d’entre nous a raison que notre désaccord n’est pas un problème. »

« Nous n’avons pas non plus besoin de savoir qui a raison pour mesurer le bruit. »

CHAPITRE 5

# Mesurer l’erreur

Il est évident qu’un biais peut provoquer des erreurs. Si votre balance affiche un poids surévalué chaque fois que vous vous pesez, si un manager enthousiaste promet toujours que ses projets vont prendre moitié moins de temps qu’ils ne demandent en réalité, ou si un commercial timoré est systématiquement pessimiste sur ses ventes futures, ces biais sont des sources d’erreurs.

Nous avons vu que le bruit produit, lui aussi, des erreurs coûteuses. Si un manager annonce tantôt que ses projets prendront moitié moins de temps qu’ils ne demandent en réalité, et tantôt qu’ils en prendront le double, il serait absurde de dire qu’il a raison « en moyenne ». Ses erreurs ne s’annulent pas : elles se cumulent.

Il est donc essentiel de mesurer les contributions respectives du biais et du bruit à l’erreur. Nous allons proposer dans ce chapitre une manière de le faire. Notre conclusion sera simple : en matière de jugements professionnels, quels qu’ils soient, et partout où la finalité du jugement est la justesse, c’est-à-dire la minimisation de l’erreur, *le biais et le bruit jouent des rôles identiques dans le calcul de l’erreur globale*. Dans certains cas, le biais pèsera plus lourd dans le calcul ; dans d’autres (bien plus fréquents qu’on ne le pense), ce sera le bruit. Mais, dans tous les cas, réduire le bruit a

le même impact sur l'erreur globale qu'une réduction du biais de même ampleur. Pour cette raison, la mesure et la réduction du bruit devraient avoir la même priorité que celles du biais.

Cette conclusion repose sur une définition particulière de l'erreur, qui a une longue histoire et qui est généralement acceptée en sciences et en statistique. Nous présenterons dans ce chapitre une rapide synthèse de cette histoire, ainsi que le raisonnement qui sous-tend cette définition de l'erreur. Les lecteurs allergiques aux formules mathématiques, et qui sont prêts à croire sur parole la conclusion que nous venons d'annoncer, pourront sans dommage se dispenser de la lecture de ce chapitre.

## *GoodSell devrait-il réduire le bruit ?*

Imaginons une entreprise de distribution du nom de GoodSell, qui emploie un grand nombre de prévisionnistes. Le travail de ces derniers consiste à prévoir les parts de marché que GoodSell obtiendra dans les différentes régions où elle opère. Amy Simkin, chef du département prévision de l'entreprise, vient justement de réaliser un audit de bruit : chacun de ses prévisionnistes a produit sa propre estimation de la part de marché future de GoodSell pour une même région.

La figure 3 montre les résultats de cet audit de bruit. Amy peut constater que la distribution des prévisions de part de marché de GoodSell obéit (avec une fidélité qu'on ne rencontre que dans les livres) à la distribution normale, appelée aussi gaussienne, et représentée par la courbe en cloche bien connue. La prévision la plus fréquente, représentée par le sommet de la cloche, est de 44 %. Amy peut également constater que le système de prévision de l'entreprise est assez sujet au bruit : dans un monde idéal, les prévisions de part de marché seraient toutes justes et donc toutes identiques, mais, en réalité, elles varient très fortement.

Pour chiffrer ce niveau de bruit dans le système de prévision de GoodSell, la mesure appropriée, comme dans l'exemple des intervalles de temps mesurés sur votre smartphone, est l'*écart-type* des prévisions. Ainsi que son nom l'indique, l'écart-type représente une distance « typique » par rapport à la moyenne. Dans cet exemple, il est de 10 points de pourcentage. Puisque nous avons affaire à une distribution normale, nous savons qu'environ deux tiers des prévisions sont comprises dans un intervalle de plus ou moins une fois l'écart-type autour de la moyenne : ici, donc, les deux tiers des prévisionnistes ont indiqué un chiffre compris entre 34 et 54 %. Amy dispose désormais d'une estimation du volume de bruit systémique dans les prévisions de parts de marché de l'entreprise. (Pour en avoir une estimation plus fiable, il faudrait un audit de bruit plus approfondi, par exemple sur plusieurs régions ou plusieurs périodes ; mais un seul cas suffit à illustrer notre propos.)

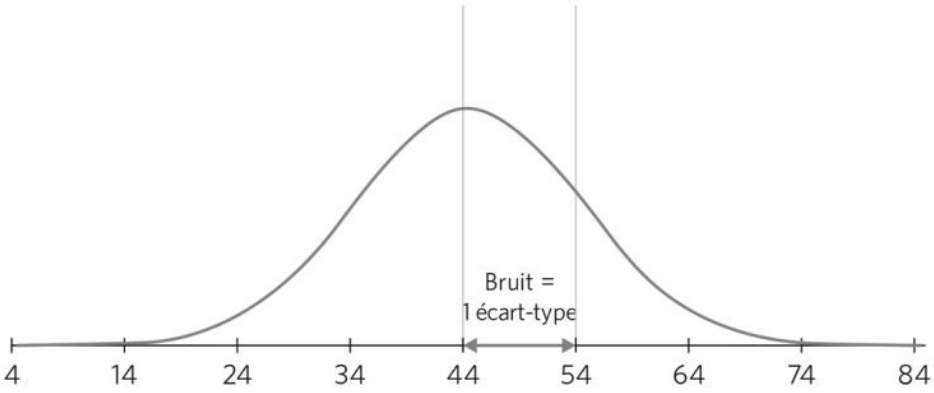



**Figure 3.** Distribution des prévisions de parts de marché de GoodSell pour une région.

Comme les dirigeants de la compagnie d'assurances du chapitre 2, Amy trouve ces résultats préoccupants et souhaite y mettre bon ordre. Le volume inacceptable de bruit suggère un manque total de rigueur dans l'application des méthodes et procédures que les prévisionnistes sont censés suivre. Amy

propose donc à son patron de faire intervenir un spécialiste du bruit. Celui-ci se fait fort de codifier les méthodes de travail des prévisionnistes afin de les rendre plus uniformes. Malheureusement, la demande d'Amy se heurte à une fin de non-recevoir. L'objection de son patron paraît d'ailleurs raisonnable : comment pourrions-nous nous attaquer aux erreurs dans nos prévisions, lui demande-t-il, tant que nous ne savons pas si ces prévisions sont justes ? À l'évidence, continue-t-il, si l'erreur moyenne des prévisionnistes est importante (c'est-à-dire s'il y a un biais important dans les prévisions), la priorité devrait être d'y remédier. Par conséquent, conclut le patron d'Amy, avant de se demander comment améliorer ses prévisions de parts de marché, GoodSell doit d'abord attendre de voir si elles s'avèrent, en moyenne, correctes.

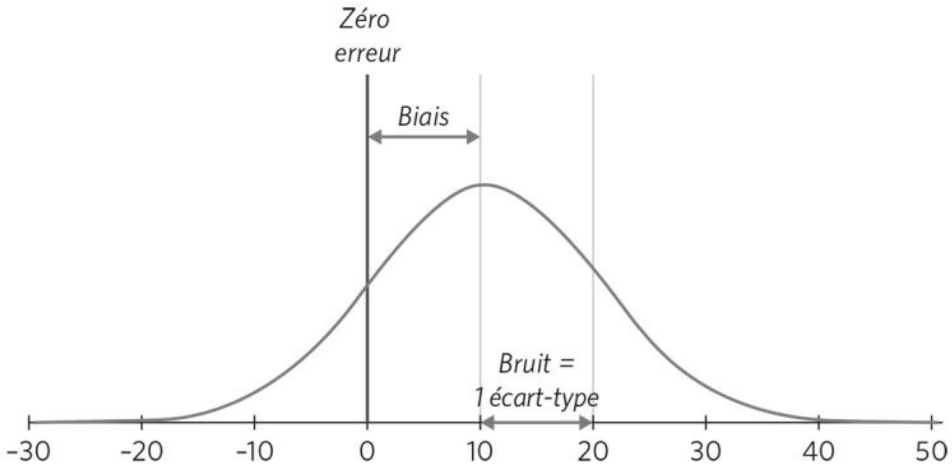



**Figure 4.** Distribution des erreurs de prévision de parts de marché de GoodSell pour une région.

Amy prend donc son mal en patience ; et un an plus tard, la réponse à cette question est connue : la part de marché réellement atteinte dans la région ciblée est de 34 %. Désormais, Amy et son patron savent quantifier l'erreur, qui est la différence entre la prévision et le résultat : pour la

prévision moyenne de 44 %, l'erreur moyenne est de + 10 points de pourcentage. La valeur du biais est donc ici égale à celle du bruit. Cette égalité n'est bien sûr pas une règle générale, mais une coïncidence que nous provoquons ici pour faciliter l'exposé.

Le même calcul d'erreur peut être fait pour chacun des prévisionnistes : l'erreur est de + 10 pour le prévisionniste dont la prévision se trouvait être égale à la moyenne de ses pairs ; de 0 pour celui qui avait fait une prévision de 34 % ; de – 10 pour celui qui avait prévu 24 %, et ainsi de suite. La figure 4 montre la distribution des *erreurs* ainsi calculées. La courbe est identique à la distribution des *prévisions* de la figure 3, à ceci près que la valeur juste (34 %) a été soustraite de chaque prévision. Mais la forme de la courbe n'a pas changé, et l'écart-type (notre mesure du bruit) est toujours de 10 points de pourcentage.

Les figures 3 et 4 se ressemblent beaucoup, mais il y a entre elles une différence importante. Cette différence est analogue à celle entre les figures 1 et 2 de l'introduction : le motif dessiné par les points d'impact sur la cible d'une part, le motif observé au dos de la cible d'autre part. Connaître la position de la cible n'était pas nécessaire pour observer le bruit affectant les tirs ; de même, connaître le résultat ne change rien à ce que l'on savait déjà du bruit dans les prévisions de parts de marché de GoodSell.

Ce qu'Amy et son patron ignoraient auparavant, et ce qu'ils connaissent désormais, c'est le niveau de biais : dans cet exemple, le biais et le bruit sont numériquement identiques et égaux à 10 points de pourcentage. Nous pouvons observer que la plupart des prévisionnistes avaient fait une erreur optimiste, c'est-à-dire qu'ils avaient surestimé la part de marché qui serait atteinte. Ainsi, la plupart des prévisions se situent à droite de la barre verticale qui dénote « zéro erreur ». Plus précisément, les propriétés de la distribution normale impliquent que 84 % des prévisions sont dans ce cas.

Comme l'observe le patron d'Amy avec une satisfaction à peine dissimulée, il avait raison. Il y avait beaucoup de biais dans les prévisions ! Et il est maintenant évident que réduire ce biais serait une bonne chose. Mais Amy se demande toujours s'il aurait été intéressant de réduire le bruit il y a un an – et si ce serait encore une bonne idée aujourd'hui. Qu'apporterait ce type d'amélioration par rapport à celle qu'apporte la réduction du biais ?

## *Les moindres carrés*

Pour répondre à la question d'Amy, nous avons besoin d'une « règle de calcul » des erreurs qui nous permette de combiner les erreurs individuelles dans une seule et même valeur de l'erreur globale. Cette règle existe : c'est la *méthode des moindres carrés*, inventée en 1795 par Carl Friedrich Gauss, un prodige des mathématiques né en 1777, et qui fit ses premières découvertes alors qu'il était encore adolescent[1].

Gauss proposa une règle nouvelle pour évaluer la contribution des erreurs individuelles à l'erreur d'ensemble. Sa mesure de l'erreur globale est appelée *erreur quadratique moyenne* et généralement abrégée *MSE*, d'après l'anglais *mean squared error*. C'est la *moyenne des carrés des erreurs* de mesure.

Les arguments de Gauss pour choisir cette manière de mesurer l'erreur globale sortent bien sûr du champ de ce livre. Mais on peut noter que sa solution n'a rien d'une évidence. Pourquoi utiliser les *carrés* des erreurs ? L'idée semble arbitraire et même étrange. Et pourtant, comme vous allez le voir, elle repose sur une intuition que vous partagez certainement.

Pour l'expliquer, faisons un détour par un problème qui va vous sembler tout à fait différent, mais qui est en réalité le même. Imaginez que l'on vous a donné une règle et que l'on vous a demandé de mesurer la longueur d'une

ligne au millimètre près. Vous avez fait cinq mesures. Elles sont représentées sur la figure 5 par les cinq triangles pointés vers le bas.

Comme vous pouvez le voir, les cinq mesures sont toutes comprises entre 971 et 980 millimètres. Quelle est la meilleure estimation possible de la longueur véritable de la ligne ? Deux réponses viennent immédiatement à l'esprit. La première possibilité est la mesure médiane, c'est-à-dire la troisième des cinq mesures quand on les classe de la plus courte à la plus longue : elle est de 973 millimètres. La seconde est la moyenne arithmétique (qu'on appelle simplement moyenne en langage courant) : elle est dans cet exemple de 975 millimètres, et indiquée sur la figure 5 par le triangle pointé vers le haut. Intuitivement, vous penchez sans doute pour la moyenne. Vous avez raison : la moyenne contient davantage d'informations. Elle tient compte de la valeur des nombres, alors que la médiane n'est affectée que par leur ordre.

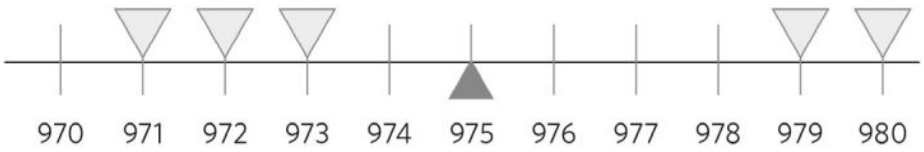



**Figure 5.** Cinq mesures de la même longueur.

Il y a un lien étroit entre ce problème d'estimation, sur lequel tout le monde a généralement une intuition assez claire, et le problème qui nous occupe ici, celui de la mesure de l'erreur globale. Il s'agit même, en réalité, des deux faces de la même pièce. Pour quelle raison ? Parce que *la meilleure estimation est celle qui minimise l'erreur globale des mesures disponibles*. De ce fait, si notre intuition est correcte quand nous pensons que la moyenne est la meilleure estimation possible, alors nous devons utiliser, pour mesurer l'erreur globale, une formule compatible avec cette intuition. En d'autres termes, l'erreur globale calculée par notre formule

doit être à son minimum quand nous prenons, comme estimation de la valeur vraie, la moyenne arithmétique.

L'erreur quadratique moyenne (MSE) possède cette propriété, et c'est la seule définition de l'erreur globale qui la possède. Dans la figure 6, nous avons calculé la valeur de la MSE dans notre ensemble de cinq mesures pour dix valeurs entières possibles de la longueur de la ligne. Par exemple, si la valeur juste était 971 millimètres, les erreurs des cinq mesures seraient de 0, 1, 2, 8 et 9. Les carrés de ces erreurs s'ajoutent pour faire 150 (0 + 1 + 4 + 64 + 81) et leur moyenne est 30 (150 divisé par 5). Cette erreur globale est élevée, car votre estimation est loin du compte. Vous pouvez voir que la MSE diminue à mesure qu'on se rapproche de 975 millimètres – la moyenne – et augmente à mesure qu'on la dépasse et qu'on s'en éloigne. La moyenne, notre estimation intuitive de la valeur juste, est bien la valeur qui minimise l'erreur globale calculée par la MSE.

Vous pouvez également constater que l'erreur globale augmente rapidement lorsque votre estimation s'écarte de la moyenne. Il suffit par exemple que votre estimation augmente de seulement 3 millimètres, passant de 976 à 979 millimètres, pour que la MSE double. C'est là une des caractéristiques essentielles de la méthode : la mise au carré donne aux grandes erreurs un poids bien plus élevé qu'aux petites.

Vous comprenez maintenant pourquoi la formule de Gauss pour mesurer l'erreur globale est l'erreur quadratique moyenne, et pourquoi sa méthode d'estimation est appelée la méthode des *moindres carrés*. La mise au carré des erreurs en est l'idée centrale, et toute autre formule serait incompatible avec l'intuition que la moyenne est la meilleure estimation.

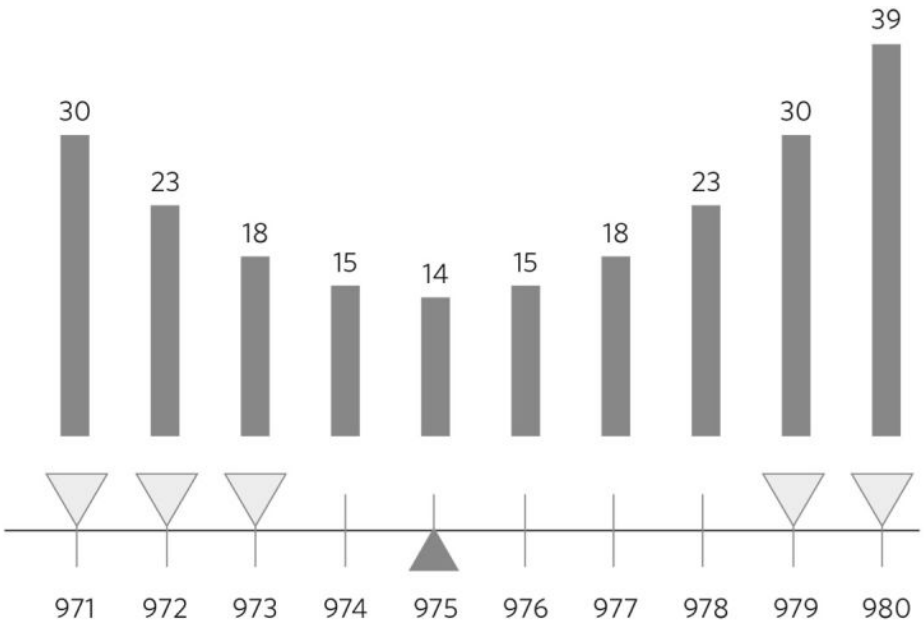



**Figure 6.** L’erreur quadratique moyenne (*mean squared error* ou MSE) pour dix estimations possibles de la longueur juste.

Les avantages de la méthode de Gauss ont été rapidement reconnus par les mathématiciens. Entre autres prouesses, Gauss s’est servi de la MSE (et de quelques autres innovations mathématiques) pour résoudre une énigme qui avait donné du fil à retordre aux plus grands astronomes européens : la redécouverte de Cérès, un astéroïde qui avait été aperçu brièvement avant de disparaître dans l’éclat du Soleil, en 1801. Les astronomes avaient essayé d’estimer la trajectoire de Cérès, mais leur méthode pour prendre en compte l’erreur de mesure de leurs télescopes était erronée, et la planète ne réapparut dans aucun des endroits du ciel qu’ils suggéraient. Gauss refit leurs calculs avec la méthode des moindres carrés. Et quand les astronomes pointèrent leurs télescopes sur le point qu’il indiquait, Cérès les y attendait !

Dans toutes les disciplines, les scientifiques ne tardèrent pas à adopter la méthode des moindres carrés. Deux siècles plus tard, elle reste la méthode standard pour évaluer les erreurs, partout où l’on cherche à optimiser l’exactitude d’une mesure. L’utilisation du carré des erreurs est au cœur des

statistiques ; et l'erreur quadratique moyenne (MSE) est l'arbitre des performances dans d'innombrables problèmes scientifiques. Comme nous allons le voir, la méthode a des implications pour le moins étonnantes.

## *L'équation de l'erreur*

Le rôle du biais et du bruit peut être résumé de manière simple par deux formules. La première de ces formules décompose l'erreur d'une seule mesure (par exemple, dans l'exemple précédent, une des cinq mesures de la longueur de la ligne). L'erreur comprend deux éléments qui nous sont désormais familiers : le biais – c'est-à-dire l'erreur moyenne – et une erreur résiduelle, propre à chaque mesure, qu'on peut appeler « erreur de bruit ». Cette erreur de bruit est positive quand l'erreur de la mesure est plus grande que le biais, négative quand elle est plus petite. La moyenne des erreurs de bruit est de zéro. Jusque-là, rien de nouveau.

**Erreur d'une seule mesure = biais + erreur de bruit**

La seconde formule, que nous appellerons *équation de l'erreur*, est plus importante : elle décompose l'erreur quadratique moyenne (MSE), la mesure de l'erreur globale que nous venons de présenter. Étant entendu que le biais est toujours l'erreur moyenne et que le bruit est l'écart-type des mesures (qui est identique à l'écart-type des erreurs de bruit), quelques calculs simples permettent de montrer que la MSE est égale à la *somme des carrés du biais et du bruit*[2] :

**Erreur globale (MSE) = Biais$^2$ + Bruit$^2$**

La forme de cette équation – une somme de deux carrés – vous rappelle peut-être un des musts de vos années de collège, le théorème de Pythagore. Souvenez-vous : dans un triangle rectangle, la somme des carrés des deux côtés les plus courts est égale au carré de l'hypoténuse (le côté le plus long). Cela permet une visualisation simple de l'équation de l'erreur, dans laquelle la MSE, le carré du bruit et le carré du biais sont les surfaces des trois carrés bordant les côtés d'un triangle rectangle. La figure 7 montre que la MSE (la surface du carré le plus sombre) est égale à la somme des surfaces des deux autres carrés. Dans la partie gauche de la figure, il y a plus de bruit que de biais ; dans celle de droite, plus de biais que de bruit. Mais la MSE est la même, et l'équation de l'erreur est vérifiée dans les deux cas.

Comme la formule mathématique et sa représentation visuelle le suggèrent toutes les deux, *le biais et le bruit jouent des rôles identiques dans l'équation de l'erreur*. Le biais et le bruit sont indépendants l'un de l'autre, et contribuent de la même manière à la détermination de l'erreur globale.

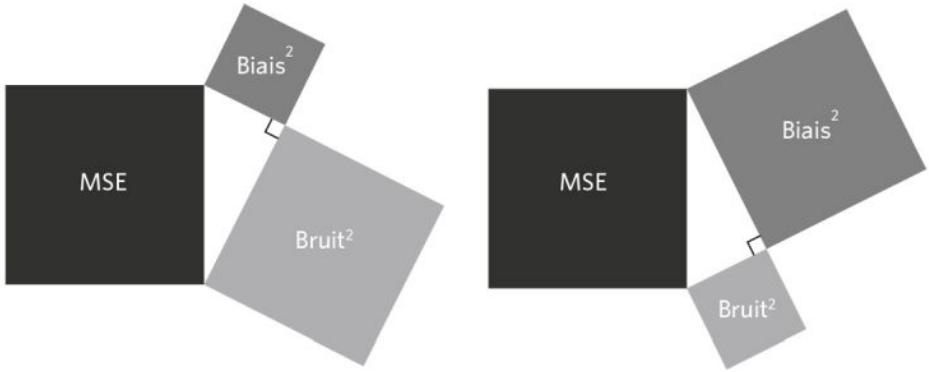



**Figure 7.** Deux décompositions de l'erreur quadratique moyenne (MSE).

L'équation de l'erreur apporte la réponse que nous cherchions à la question pratique posée par Amy : vaut-il mieux, pour réduire l'erreur globale, réduire le biais ou réduire le bruit ? La réponse est simple : comme

le biais et le bruit sont interchangeables dans l'équation de l'erreur, on obtiendra un résultat strictement identique en diminuant l'un ou l'autre d'un même montant. Dans la figure 4, il se trouve que le biais et le bruit sont égaux (10 points de pourcentage tous les deux), et leurs contributions à l'erreur globale sont égales.

L'équation de l'erreur apporte aussi un appui sans équivoque à l'élan initial d'Amy pour réduire le bruit. Chaque fois qu'on détecte du bruit, on devrait s'efforcer de le réduire ! Et il est maintenant clair que le patron d'Amy se trompait quand il soutenait que GoodSell devait d'abord mesurer le biais dans ses prévisions, avant de décider des mesures à prendre. En ce qui concerne l'erreur globale, le biais et le bruit sont indépendants : quel que soit le niveau de biais, le gain résultant de la réduction du bruit est le même.

Cette idée essentielle peut s'avérer extrêmement contre-intuitive. Pour l'illustrer, la figure 8 montre l'effet d'une réduction équivalente du biais et du bruit. Pour vous aider à visualiser la comparaison avant-après dans chacun des deux panels, la distribution initiale des erreurs (celles de la figure 4) est représentée par une courbe en pointillé.

Dans le panel A, nous supposons que le patron d'Amy a décidé de faire les choses à sa façon : il a réussi à évaluer le biais, puis à le réduire de moitié (peut-être en donnant du feed-back aux prévisionnistes trop optimistes). Rien n'a été fait pour le bruit. L'amélioration saute aux yeux : toute la distribution des prévisions de parts de marché s'est déplacée pour se rapprocher de la valeur juste.

Dans le panel B, nous voyons ce qui serait passé si Amy avait mis en œuvre son plan de réduction du bruit. Le biais reste inchangé, mais le bruit a été réduit de moitié. Le paradoxe, c'est qu'à première vue, cette baisse du niveau de bruit semble avoir aggravé les choses. Les prévisions de parts de marché sont désormais plus concentrées (et pour cause : il y a moins de bruit), mais elles ne sont pas plus justes en moyenne (puisqu'il n'y a pas

moins de biais). Autrement dit, alors que 84 % des prévisionnistes se trompaient dans le même sens, celui de l'optimisme ou de la surestimation de la part de marché, presque tous (98 %) sont désormais dans ce cas. La réduction du bruit semble avoir rendu les prévisions *plus précisément fausses* : ce n'est pas vraiment l'amélioration qu'espérait Amy !

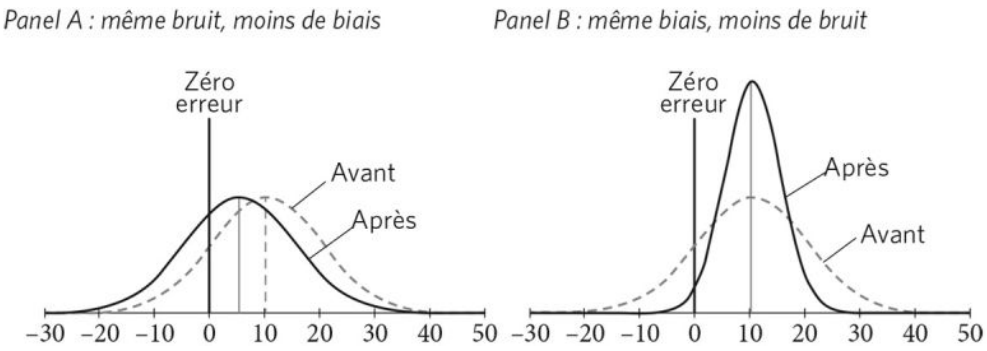



**Figure 8.** Distribution des erreurs quand on réduit de moitié soit le biais, soit le bruit.

En dépit des apparences, cependant, l'erreur globale a bien été réduite d'autant dans le panel B que dans le panel A. Le sentiment de détérioration que nous éprouvons dans le panel B trouve sa source dans une intuition répandue, mais erronée. La bonne mesure du biais n'est pas la proportion des erreurs positives et négatives : c'est l'erreur moyenne. Dans la figure 8, c'est la distance entre le sommet de la courbe en cloche et la valeur juste (« zéro erreur »). Regardez le panel B : entre « avant » et « après », cette erreur moyenne n'a pas changé. Elle est toujours de 10 %. Il est vrai que la présence du biais est désormais plus frappante, car il représente une proportion plus importante de l'erreur globale (80 % au lieu de 50 %) ; mais c'est seulement parce que le bruit a été réduit. Inversement, dans le panel A, le biais a été réduit, mais pas le bruit. Le résultat net est que la MSE est la même dans les deux panels : encore une fois, réduire le bruit ou réduire le biais de la même quantité a le même effet sur l'erreur globale.

Comme cet exemple le montre, l'utilisation de la MSE contredit beaucoup de nos intuitions sur l'évaluation des jugements prédictifs. Pour minimiser la MSE, il faut surtout éviter les grandes erreurs. Quand on mesure une longueur, par exemple, passer d'une erreur de 11 centimètres à une erreur de 10 centimètres réduit la MSE de 121 à 100. C'est là une amélioration vingt et une fois plus importante que lorsqu'on réduit à zéro une erreur de 1 centimètre pour tomber parfaitement juste. Malheureusement, nos intuitions à cet égard sont presque à l'exact opposé de ce qu'elles devraient être[3] : nous voulons absolument mettre dans le mille, et nous sommes extrêmement sensibles aux petites erreurs, mais nous nous soucions assez peu de la différence entre deux grandes erreurs. Ainsi, même quand nous nous efforçons sincèrement de produire des jugements exacts, notre réaction émotionnelle à la taille des erreurs peut entrer en conflit avec la définition scientifique de l'exactitude.

Il va sans dire que la meilleure solution pour GoodSell – et en général – serait de réduire à la fois le biais et le bruit. Comme le biais et le bruit sont indépendants l'un de l'autre, il n'y a pas de raison de choisir entre Amy et son patron. On peut d'ailleurs noter que, si GoodSell décide de réduire le bruit, le fait que cette réduction rende le biais plus visible apporte un avantage annexe : sans bruit, le biais devient si évident qu'il est impossible de l'ignorer (encore une fois, regardez le panel B). Ainsi, en réduisant le bruit, l'entreprise s'assure que la réduction du biais sera sa priorité suivante.

Reste une situation dans laquelle la réduction du bruit ne devrait pas être la première priorité : quand le biais est largement supérieur au bruit. Mais, à ce propos, on peut tirer de l'exemple de GoodSell un autre enseignement important. Dans ce modèle simplifié, nous avons supposé que le bruit et le biais étaient égaux. Leurs contributions respectives à l'erreur globale sont elles aussi égales : avant les interventions d'Amy ou de son patron, le biais et le bruit expliquent chacun 50 % de l'erreur globale. Comme nous l'avons noté, cependant, 84 % des prévisionnistes font une

erreur dans la même direction. Il faut donc, pour que le biais produise autant d'erreurs que le bruit, un biais considérable : six personnes sur sept se trompant dans le même sens ! Un tel biais a peu de chances de passer inaperçu. Nous ne devrions donc pas nous étonner de trouver bien des situations dans lesquelles il y a, en réalité, plus de bruit que de biais.

Une précision technique : nous avons illustré l'application de l'équation de l'erreur à un cas unique, une seule des régions où opère GoodSell. Bien sûr, il est toujours souhaitable de réaliser un audit de bruit sur plusieurs cas en même temps. Le raisonnement est exactement le même. On calcule l'équation de l'erreur séparément pour chacun des cas, et on obtient une équation globale en prenant les moyennes de la MSE, du biais au carré et du bruit au carré pour l'ensemble des cas. Si Amy avait suivi cette démarche, et demandé des prévisions de parts de marché multiples pour plusieurs régions, faites ou non par les mêmes prévisionnistes, elle aurait eu une idée plus précise du biais et du bruit dans le système de prévision de GoodSell.

## *Le coût du bruit*

L'équation de l'erreur est le fondement intellectuel de ce livre. C'est elle qui justifie l'objectif de réduction du bruit systémique dans les jugements prédictifs, et qui explique pourquoi cet objectif devrait, en principe, être aussi important que celui de réduction du biais statistique. (Pour éviter tout malentendu, soulignons ici – nous y reviendrons – que le biais statistique n'est pas synonyme de discrimination sociale, mais simplement une mesure de l'erreur moyenne dans un ensemble de jugements.)

Comme nous l'avons noté, l'équation de l'erreur et les conclusions que nous en tirons reposent sur l'utilisation de l'erreur quadratique moyenne (MSE) comme mesure de l'erreur globale. La règle vaut pour tous les jugements purement prédictifs, comme les prévisions et les estimations de

toutes sortes, puisque leur but est bien d'approcher d'une valeur juste ou vraie avec le maximum de précision (le moins de bruit possible) et d'exactitude (le moins de biais possible).

Néanmoins, l'équation de l'erreur ne s'applique pas aux jugements évaluatifs. D'abord, le concept d'erreur repose sur l'existence d'une valeur juste : il est difficile à appliquer dans ce type de jugement, où la valeur juste n'est pas aisément définissable. De surcroît, même si les erreurs pouvaient être spécifiées, leurs coûts ne seraient généralement pas symétriques, et il n'y a guère de raisons de penser qu'ils s'accroîtraient en proportion exacte de leur carré.

Considérez par exemple le cas d'une entreprise d'ascenseurs qui doit estimer la charge maximale d'un ascenseur. Les conséquences des erreurs sont évidemment asymétriques. Sous-estimer la charge utile est coûteux, puisque les spécifications commerciales de l'ascenseur seront réduites. Mais la surestimer peut entraîner un accident catastrophique. De même, quand la compagnie d'assurances du chapitre 2 fixe le prix d'une police d'assurance ou estime la valeur d'un sinistre, toutes les erreurs sont coûteuses, mais il n'y a pas de raison de penser que les coûts des erreurs dans les deux sens soient équivalents.

L'erreur quadratique moyenne n'est pas non plus pertinente quand il s'agit de décider du moment de partir de chez vous pour prendre un train : que vous soyez en retard d'une minute ou de cinq, vous raterez votre train de la même manière. Mettre les erreurs au carré ne correspond à aucune réalité pratique, puisqu'un retard de cinq minutes ne vous coûtera pas vingt-cinq fois plus cher qu'un retard d'une minute.

Ces exemples soulignent la nécessité de préciser les rôles respectifs des jugements prédictifs et des jugements évaluatifs dans les décisions. Une des maximes largement acceptées de la bonne prise de décision est de ne pas mélanger les faits dont vous disposez et les valeurs qui président à la décision. La bonne prise de décision doit d'abord reposer sur des jugements

*prédictifs* aussi objectifs et corrects que possible. Vous avez sans doute des préférences, des valeurs, des espoirs, des craintes, mais rien de tout cela ne doit, dans un premier temps, entrer en ligne de compte.

Ainsi, pour l'entreprise d'ascenseurs, la première étape sera de calculer de façon aussi neutre que possible la charge technique maximale de l'ascenseur en fonction des données disponibles. Il s'agit là d'un jugement objectif et prédictif, pour lequel la minimisation de la MSE est bien la mesure pertinente. Ce n'est qu'ensuite qu'intervient un second jugement, cette fois évaluatif, qui détermine le choix de la marge de sécurité nécessaire pour fixer la charge maximale admissible, celle qui sera affichée dans la cabine. Nul doute que la sécurité soit alors une considération primordiale (informée par d'autres éléments factuels et d'autres jugements prédictifs).

De même, pour décider quand partir prendre son train, la première étape devrait être de déterminer de manière objective les probabilités des différents temps de trajet. Les erreurs positives et négatives (rater son train d'un côté, perdre son temps à attendre à la gare de l'autre) ont bien des coûts différents ; mais ces coûts ne deviennent pertinents que dans un second temps, pour choisir le risque que l'on est prêt à accepter.

La même logique s'applique à des décisions aux conséquences autrement plus importantes. Un commandant militaire doit peser un grand nombre de considérations avant de décider de lancer ou non une offensive, mais une bonne partie des renseignements sur lesquels il s'appuie relèvent du jugement prédictif. Dans sa réaction à une pandémie, un gouvernement doit peser les avantages et les inconvénients de différentes options, mais aucune évaluation n'est possible sans des prévisions exactes sur les conséquences probables des options possibles (y compris celle qui consiste à ne rien faire).

Dans tous ces exemples, la décision finale nécessite un jugement évaluatif. Pour faire le choix optimal, les personnes qui décident doivent

considérer de multiples options et mesurer les avantages et inconvénients de celles-ci à l'aune de leurs propres valeurs. Mais toutes les décisions dépendent de prédictions sous-jacentes ; et les jugements prédictifs en question devraient toujours être neutres du point de vue des valeurs ou des préférences personnelles.

---

L'objectif des jugements prédictifs est toujours l'exactitude – toucher au plus près du cœur de la cible. Et l'erreur quadratique moyenne (MSE) est bien la mesure qui convient à ce type de jugement. Pour cette raison, et à condition bien sûr que la réduction du bruit ne se traduise pas par un accroissement plus grand du biais, tous les jugements prédictifs gagnent à ce que le bruit soit réduit.

### À propos de l'équation de l'erreur

« Nous devrions mesurer l'erreur dans nos prévisions comme le font des scientifiques : en utilisant la MSE. »
« Curieusement, que l'on réduise le biais ou le bruit, l'effet sur la justesse des jugements est le même. »
« Réduire le bruit dans les jugements prédictifs est toujours utile, quoi que vous sachiez – ou non – du biais. »
« Quand, sur 100 jugements, 84 sont au-dessus de la valeur juste et 16 au-dessous, il semble bien qu'il y ait un biais important. Mais c'est alors que le biais et le bruit sont égaux. »
« Des jugements prédictifs sont nécessaires dans toutes les décisions, et l'exactitude devrait être leur unique objectif. Vous devez veiller à séparer les faits dont vous disposez et les valeurs qui sont les vôtres. »

CHAPITRE 6

# L’analyse du bruit

Dans le précédent chapitre, nous avons analysé la variabilité dans la mesure d’une seule quantité, ou dans le jugement d’un seul cas. Lorsqu’on s’intéresse à un cas unique, toute variabilité dans le jugement est de l’erreur, et les deux composantes de cette erreur sont le biais et le bruit. Mais les organisations comme les compagnies d’assurances, ou plus généralement les systèmes comme le système judiciaire, ne traitent pas d’un cas unique. Ils sont conçus pour traiter des cas nombreux et disparates, et pour procéder entre eux à des distinctions. Les juges fédéraux et les experts en sinistres ne seraient guère utiles s’ils produisaient le même jugement dans toutes les affaires qui leur sont présentées.

Quand on traite de cas différents, les variations entre jugements sont intentionnelles. La variabilité dans les jugements d’un même cas reste, elle, indésirable : c’est du bruit systémique. Dans tout système où plusieurs personnes jugent plusieurs cas, il va donc falloir distinguer la variation intentionnelle du bruit systémique. Comme nous allons le voir, c’est l’objet d’un audit de bruit.

## *Un audit de bruit dans la détermination des peines*

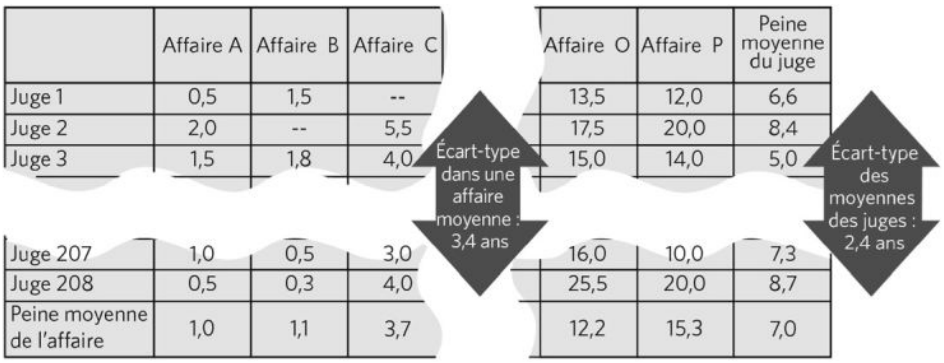
Pour illustrer l'analyse du bruit dans un système, nous allons utiliser un exemple d'audit de bruit exceptionnellement détaillé. Il concerne les peines prononcées par des juges fédéraux aux États-Unis, mais les enseignements que nous en tirerons peuvent être étendus à d'autres jugements professionnels de toutes sortes. L'étude fut publiée en 1981 dans le cadre du mouvement de réforme de la peine évoqué au chapitre 1[1]. L'objectif de cet audit de bruit était d'aller au-delà des exemples de bruit flagrants mais anecdotiques qu'avait collectés le juge Frankel. Il s'agissait donc d'évaluer de façon plus systématique « l'ampleur de la disparité des peines ».

Pour ce faire, les auteurs de l'étude eurent recours à seize descriptions simplifiées d'affaires criminelles, dans lesquelles le prévenu avait été jugé coupable et devait être condamné. Huit des cas présentaient des affaires de vol, huit des affaires d'escroquerie. Les affaires différaient sur six autres points : le prévenu pouvait être l'auteur principal du crime ou seulement un complice, il pouvait avoir ou non un casier judiciaire, il pouvait (dans le cas du vol) s'être servi ou non d'une arme, le butin pouvait être plus ou moins important, et ainsi de suite.

Les chercheurs réalisèrent des entretiens soigneusement structurés avec un échantillon national de deux cent huit juges fédéraux en activité. En quatre-vingt-dix minutes, chaque juge examinait les seize cas, puis fixait la peine qui lui semblait appropriée[2].

Pour visualiser les données ainsi collectées, imaginez un tableau de seize colonnes pour les crimes, allant de A à P, et de 208 lignes pour les juges, allant de 1 à 208. Chaque cellule, de A1 à P208, indique donc la peine de prison qu'un juge particulier fixe dans une affaire donnée. La figure 9 montre à quoi ressemblerait ce tableau de 3 328 cellules. Pour étudier le bruit, nous allons nous concentrer sur les seize colonnes : chacune

d’entre elles – chaque affaire évaluée séparément par deux cent huit juges – constitue en soi un audit de bruit.




| | Affaire A | Affaire B | Affaire C | Affaire O | Affaire P | Peine moyenne du juge |
|---|---|---|---|---|---|---|
| Juge 1 | 0,5 | 1,5 | -- | 13,5 | 12,0 | 6,6 |
| Juge 2 | 2,0 | -- | 5,5 | 17,5 | 20,0 | 8,4 |
| Juge 3 | 1,5 | 1,8 | 4,0 | 15,0 | 14,0 | 5,0 |
| Juge 207 | 1,0 | 0,5 | 3,0 | 16,0 | 10,0 | 7,3 |
| Juge 208 | 0,5 | 0,3 | 4,0 | 25,5 | 20,0 | 8,7 |
| Peine moyenne de l’affaire | 1,0 | 1,1 | 3,7 | 12,2 | 15,3 | 7,0 |

**Figure 9.** Une représentation de l’étude sur les peines.

## *La peine moyenne*

Il n’existe pas de moyen objectif de décider quelle est la « juste » peine dans une affaire donnée. Dans la suite de ce chapitre, nous ferons l’hypothèse que la moyenne des 208 peines prononcées dans chaque affaire (la *peine moyenne de l’affaire* qui apparaît dans la dernière ligne du tableau) est une approximation de la « juste peine ». Comme nous l’avons noté au chapitre 1, la Commission des peines du Congrès a eu recours au même procédé en faisant de l’historique des affaires passées le fondement de ses directives. Nous faisons donc l’hypothèse qu’il n’y a aucune erreur dans la peine moyenne de chaque affaire – autrement dit, qu’il y a zéro biais.

Nous sommes pleinement conscients que, dans le monde réel, cette hypothèse est fausse. Le jugement moyen dans certaines affaires est très probablement biaisé par rapport au jugement moyen dans d’autres affaires très semblables, par exemple en raison de la discrimination raciale qui pénalise certains prévenus. La variabilité de ces biais d’un cas à l’autre –

parfois à l'avantage du prévenu, parfois à son détriment – est une source importante d'erreur et d'injustice[3]. L'objet principal de ce chapitre – et de ce livre – est d'analyser le bruit, mais nous ne sous-estimons nullement l'importance de mesurer et de combattre aussi les erreurs moyennes, c'est-à-dire les biais. Le juge Frankel, quand il soulignait l'injustice du bruit, attirait déjà l'attention, lui aussi, sur le risque de biais (notamment lié à la discrimination raciale).

Par commodité, les affaires sont disposées par ordre croissant de gravité : la peine moyenne dans l'affaire A est de 1 an de prison, et dans l'affaire P de 15,3 ans. La peine de prison moyenne pour les seize cas est de 7 ans[4].

Imaginez maintenant un monde parfait dans lequel tous les juges se comportent comme des instruments de mesure infaillibles, des balances de la justice qui seraient totalement exemptes de bruit. À quoi ressemblerait la figure 9 dans ce monde parfait ? À l'évidence, toutes les cellules de la colonne A seraient identiques, car tous les juges prononceraient dans l'affaire A exactement la même peine : 1 an de prison. Il en irait de même dans toutes les autres colonnes. Les chiffres de chaque ligne, bien sûr, varieraient toujours, puisque chaque affaire est différente. Mais chaque ligne serait identique à celle qui suit et à celle qui précède. En d'autres termes, les différences entre les affaires seraient la seule source de variabilité dans le tableau.

Malheureusement, le monde de la justice fédérale n'est pas parfait. Les juges ne sont pas identiques, et la variabilité au sein de chaque colonne est forte, ce qui indique la présence de bruit dans les jugements de chaque affaire. Il y a plus de variabilité dans les peines qu'il ne devrait y en avoir, et le but de l'étude est de l'analyser.

## *La loterie des peines*

Partons de l'image du monde parfait que nous avons décrit ci-dessus, et où, dans chaque affaire, tous les juges prononceraient la même peine. Chaque colonne est une série de 208 chiffres identiques, reprenant la peine moyenne de l'affaire. Maintenant, ajoutez du bruit en parcourant de haut en bas chaque colonne et en changeant un chiffre ici et là : tantôt en ajoutant du temps de prison à la durée de la peine moyenne, tantôt en en retranchant. Comme les changements que vous faites ne sont pas tous identiques, ils créent de la variabilité à l'intérieur de la colonne. Cette variabilité est du bruit.

Dans le monde réel, il n'y a pas besoin d'ajouter artificiellement du bruit pour en trouver. Le résultat principal de l'audit de bruit des juges fédéraux est un niveau assourdissant de bruit dans les jugements de chaque affaire. On peut calculer notre mesure habituelle du bruit, l'écart-type, dans chaque affaire. Le résultat varie selon les affaires ; mais pour l'affaire moyenne, la peine moyenne était de 7,0 ans de prison, et l'écart-type autour de cette moyenne était de 3,4 ans[5].

Est-ce beaucoup ? Même si la notion d'*écart-type* ne vous est pas inconnue, il peut être utile d'en donner une description concrète. Imaginez que vous preniez deux juges au hasard et que vous calculiez la différence entre les jugements qu'ils prononcent dans une affaire donnée. Répétez l'opération pour toutes les paires possibles de juges, et faites la moyenne des résultats. Cette mesure (la *moyenne des valeurs absolues des différences*) devrait vous donner une idée de la loterie devant laquelle se retrouve le prévenu dans un tribunal fédéral. En supposant que les jugements soient normalement distribués, elle est de 1,128 fois l'écart-type[6]. Cela signifie que la différence moyenne entre deux peines choisies au hasard pour la même affaire sera de 3,8 ans. Au chapitre 3, nous avons évoqué la loterie à laquelle est soumis le client d'une compagnie d'assurances. Le moins qu'on puisse dire est que la loterie de la justice pénale a des conséquences plus graves.

Une différence moyenne de 3,8 ans entre des juges qui examinent la même affaire, quand la peine moyenne est de 7,0 ans, est un résultat dérangeant et, à nos yeux, inacceptable. Hélas, il y a pourtant de bonnes raisons de penser que le bruit dans l'administration réelle de la justice est encore bien supérieur. D'abord, les participants à l'audit de bruit examinaient une série d'affaires stylisées et présentées successivement, ce qui en rendait la comparaison anormalement facile. Il est bien plus aisé de rester cohérent dans un exercice de ce type que dans la vraie vie. Ensuite, les juges dans un tribunal disposent de beaucoup plus d'informations qu'ils n'en avaient dans cet audit ; et toutes ces informations sont autant de raisons pour les juges de différer les uns des autres (sauf s'il s'agit d'informations décisives). Pour ces raisons, nous pensons que le niveau de bruit auquel les prévenus sont soumis dans un vrai tribunal est plus grand encore que ne le mesure l'audit que nous avons résumé ici.

## *Des juges plus ou moins sévères : le bruit de niveau*

Les auteurs de l'étude ne se sont pas arrêtés à ce constat : dans l'étape suivante de l'analyse, ils ont décomposé le bruit en plusieurs éléments distincts. La première interprétation du bruit qui vous est sans doute venue à l'esprit – comme elle était venue à l'esprit du juge Frankel – est que le bruit provient de la disposition des magistrats à prononcer des peines plus ou moins sévères. Comme vous le diront tous les habitués des prétoires, chaque juge a sa réputation : certains passent pour des « durs », plus sévères que le juge moyen, d'autres pour des « tendres », plus cléments que la moyenne de leurs pairs. À ces écarts, nous donnons le nom d'*erreurs de niveau*. (Une fois encore, l'erreur est définie ici comme l'écart par rapport à

la moyenne, que nous considérons comme une approximation de la « juste » peine. Une « erreur » peut donc contribuer à corriger une injustice, si la peine moyenne est injuste.)

Dans tous les types de jugements, les erreurs de niveau varient d'une personne à l'autre. Dans les évaluations de performance, certains superviseurs notent plus généreusement. Quand ils font des prévisions de parts de marché, certains prévisionnistes sont plus optimistes. Et quand il s'agit de recommander ou non une opération chirurgicale à des patients qui souffrent du dos, certains médecins se montrent plus interventionnistes et recommandent l'opération à un pourcentage plus élevé de leurs patients.

Dans le tableau de la figure 9, chaque ligne correspond aux peines prononcées par un même juge, et la moyenne des peines que prononce celui-ci dans les seize affaires différentes apparaît dans la colonne la plus à droite. Cette *peine moyenne du juge* est une mesure de son niveau de sévérité. Et en examinant cette colonne, on découvre que les juges varient fortement dans leurs niveaux de sévérité. L'écart-type des valeurs dans cette colonne est de 2,4 ans. Cette variabilité n'a strictement rien à voir avec la justice, et tout à voir avec le juge. Les différences de peine moyenne reflètent les caractéristiques de chaque juge : ses origines, son parcours de vie, ses opinions politiques, ses biais, et ainsi de suite.

Pour comprendre l'origine de ces différences, les auteurs de l'étude ont notamment demandé aux juges de leur indiquer ce qu'ils considèrent comme la finalité principale de la peine pénale en général : vise-t-elle à mettre hors d'état de nuire le criminel (en le retirant de la société) ? À le réhabiliter ? Ou à dissuader d'autres délinquants potentiels ? Sans surprise, les juges pour lesquels la réhabilitation est la principale finalité de la peine avaient tendance à prononcer des peines de prison plus courtes (et à choisir plus souvent des peines de substitution), tandis que ceux pour qui la dissuasion ou la mise hors d'état de nuire sont primordiales se montraient plus sévères. Par ailleurs, les juges basés dans les États du sud des États-

Unis prononçaient des peines de prison significativement plus longues que leurs homologues d'autres régions. Enfin, bien sûr, la sévérité moyenne des peines était plus grande chez les juges politiquement conservateurs.

La conclusion générale est que le niveau moyen des condamnations fonctionne peu ou prou comme un trait de personnalité. Vous pourriez vous servir de cette étude pour placer les juges sur une échelle allant de « très sévère » à « très clément », de la même façon qu'un test de personnalité mesure leur degré d'extraversion ou de sociabilité. Comme d'autres traits de caractère, la sévérité de la peine est peut-être corrélée à des facteurs génétiques, aux parcours de vie des individus en question et à d'autres aspects de leur personnalité. Une chose est certaine : rien de tout cela n'a le moindre rapport avec l'affaire ou avec le prévenu. Nous parlons de *bruit de niveau* pour désigner cette variabilité dans les jugements moyens des juges.

## *Des juges qui diffèrent dans leur appréciation de chaque affaire : le bruit de pattern*

Comme le montrent les flèches noires sur le tableau de la figure 9, le bruit de niveau est de 2,4 ans et le bruit systémique de 3,4 ans. Cette différence indique que les différences de sévérité moyenne entre les juges (le bruit de niveau) ne suffisent pas à expliquer le bruit systémique. Le bruit a donc une autre composante, que nous appellerons le *bruit de pattern*.

Pour comprendre le bruit de pattern, revenons à la figure 9 et prenons une cellule au hasard – par exemple la cellule C3. La peine moyenne pour l'affaire C figure au bas de la colonne : comme vous pouvez le voir, elle est de 3,7 ans. Allons maintenant à la colonne la plus à droite pour trouver la peine moyenne donnée par le juge 3 pour les différentes affaires : elle est de

5,0 ans, soit 2,0 ans de moins que la moyenne générale. Si la variation de la sévérité des juges était la seule source de bruit dans la colonne 3, on devrait s'attendre à ce que la peine prononcée en C3 soit la peine moyenne de l'affaire corrigée de la différence de sévérité du juge : 3,7 – 2,0 = 1,7 an. Or le chiffre donné en C3 est de 4,0 ans. Apparemment, le juge 3 – bien qu'il soit d'ordinaire plus clément que la moyenne de ses pairs – a prononcé, *dans cette affaire*, une peine particulièrement sévère.

On pourrait chercher à appliquer la même logique additive à tout le tableau[7], et, dans la plupart des cas, on trouverait, comme dans l'exemple de la cellule C3, des écarts entre ce calcul et la réalité. Si vous regardez une ligne entière, vous verrez que les juges ne sont pas aussi sévères dans toutes les affaires : ils le sont plus que leur moyenne personnelle pour certaines d'entre elles et moins pour d'autres. Chacun a un motif propre, un *pattern* de jugement individuel. Pour cette raison, nous appelons ces écarts résiduels des *erreurs de pattern*. Si vous calculiez l'erreur de pattern pour chaque cellule du tableau, vous verriez que leur somme algébrique est égale à zéro pour tous les juges (en ligne) et pour toutes les affaires (en colonne). Mais les erreurs de pattern ne s'annulent pas les unes les autres dans leur contribution au bruit, car les valeurs de chaque cellule sont élevées au carré pour le calcul de celui-ci.

Il y a un moyen plus simple de confirmer que le simple modèle additif de la peine ne tient pas, et que les erreurs de niveau ne suffisent pas à expliquer le bruit. Vous pouvez voir dans la dernière ligne du tableau que les *peines moyennes de l'affaire* augmentent progressivement quand on se déplace de gauche à droite : c'est l'ordre dans lequel nous les avons classées. Mais il n'en va pas de même au niveau des lignes. Le juge 208, par exemple, a prononcé une peine beaucoup plus sévère dans l'affaire O que dans l'affaire P. Quand les juges rendent leurs sentences dans des affaires différentes, ils établissent, implicitement, un classement de ces

affaires par ordre de sévérité. Et leurs classements ne sont pas identiques – loin de là.

Nous utilisons l'expression *bruit de pattern* pour désigner la variabilité que nous venons d'identifier. Cette variabilité reflète un pattern complexe dans les attitudes des juges par rapport aux affaires qu'ils jugent. Ainsi, un juge peut être plus sévère que la moyenne de manière générale, mais plus clément (toujours par rapport à la moyenne) pour les criminels en col blanc. Un autre peut avoir tendance à l'indulgence, mais punir plus sévèrement les récidivistes. Un troisième peut être proche de la sévérité moyenne, mais inhabituellement compréhensif quand le prévenu n'est qu'un complice, et inhabituellement dur quand la victime est une personne âgée. (Un statisticien dirait qu'il existe, en plus de l'« effet juge » et de l'« effet affaire », des *interactions* entre juges et affaires.)

S'agissant de la justice criminelle, certaines réactions idiosyncratiques peuvent correspondre à la philosophie personnelle du juge en matière de peine. D'autres réactions peuvent être le fruit d'associations d'idées dont le juge est à peine conscient : un prévenu peut lui rappeler un criminel particulièrement haïssable, un autre peut au contraire lui faire penser à son propre fils. Comme le montrent ces exemples, ces patterns ne sont pas aléatoires : on pourrait s'attendre à ce qu'ils réapparaissent si le juge traitait la même affaire une deuxième fois. Mais puisque le bruit de pattern est, en pratique, très difficile à prévoir, il ajoute encore de l'incertitude à la loterie déjà imprévisible de la peine. Comme le notent les auteurs de l'étude, « les différences de pattern entre les juges dans l'influence des caractéristiques du délit et du délinquant [sont] une forme supplémentaire de disparité de la peine[8] ».

Peut-être avez-vous noté que la décomposition du bruit systémique en bruit de niveau et bruit de pattern suit la même logique que l'équation de l'erreur introduite au chapitre précédent, lorsque nous décomposions

l’erreur en biais et en bruit. Cette fois, l’équation peut être rédigée de la façon suivante :

$$\textbf{(Bruit systémique)}^2 = \textbf{(bruit de niveau)}^2 + \textbf{(bruit de pattern)}^2$$

Comme l’équation de l’erreur, cette équation peut être représentée visuellement en utilisant le théorème de Pythagore (figure 10). Nous avons représenté ici deux des côtés du triangle comme égaux. La raison en est que, dans l’étude sur les peines, le bruit de pattern et le bruit de niveau contribuent au bruit systémique à parts sensiblement égales[9].

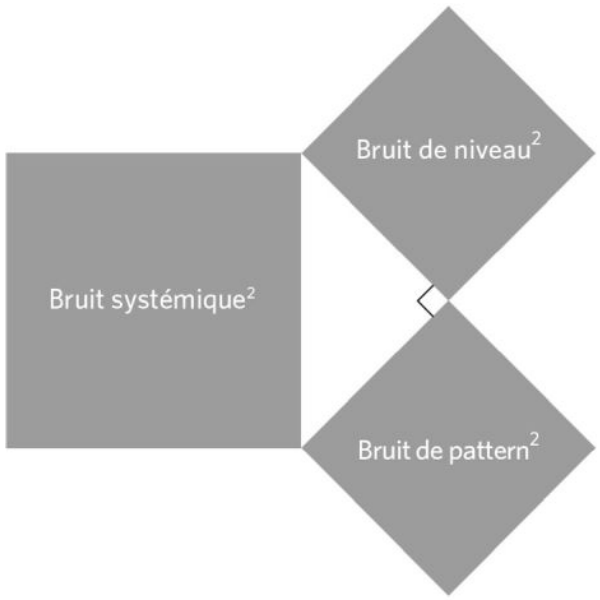



**Figure 10.** Décomposer le bruit systémique.

Le bruit de pattern est omniprésent. Quand des médecins décident ou non d’hospitaliser des patients, quand des entreprises choisissent d’embaucher tel ou tel candidat, quand des avocats décident de porter telle ou telle affaire devant la justice, ou quand des producteurs décident de produire telle ou telle émission de télévision, la hiérarchie qu’ils établissent

entre les patients, les candidats, les affaires ou les émissions varie d'une personne à l'autre. C'est l'effet du bruit de pattern.

## *Les composantes du bruit*

Nous avons jusqu'ici décrit le bruit de pattern comme résultant des caractéristiques personnelles de chaque juge. Ce faisant, nous avons laissé de côté une question importante : la possible contribution de l'erreur aléatoire.

Rappelez-vous l'exercice du chronomètre : quand vous avez essayé de mesurer plusieurs fois de suite un intervalle de temps de 10 secondes, vos résultats ont varié d'une mesure à l'autre. Cela témoignait d'une variabilité intra-individuelle. Pour la même raison, les juges n'auraient pas prononcé précisément les mêmes peines pour les seize affaires si on leur avait demandé de le faire plusieurs fois de suite. D'ailleurs, comme nous le verrons, ils n'auraient pas prononcé non plus les mêmes peines si l'étude avait été réalisée un autre jour de la même semaine. Quand un juge est de bonne humeur parce qu'il fait beau, qu'il a reçu une bonne nouvelle de sa fille ou que son équipe de football favorite a gagné la veille, il peut se montrer plus clément qu'il ne l'aurait été sans cela.

Nous appellerons *bruit occasionnel* cette variabilité entre occasions de jugement due à des effets transitoires. Conceptuellement, cette variabilité est distincte des différences stables entre personnes que nous venons de décrire, qui sont des variations interindividuelles. Mais, en pratique, les deux sont confondues : le bruit de pattern que nous avons observé comprend l'une et l'autre.

Pourquoi, alors, avons-nous, dans l'analyse des peines, mis l'accent sur les variations interindividuelles ? Pourquoi avons-nous choisi de considérer que le bruit de pattern correspond à des attitudes stables chez les juges,

plutôt qu'à du bruit occasionnel ? Peut-être sommes-nous trop optimistes, mais nous pensons que le bruit occasionnel n'a joué, dans ce cas particulier, qu'un rôle mineur. Les juges expérimentés qui ont participé à l'étude avaient certainement des idées préétablies sur le poids à accorder à chaque aspect d'une affaire, et c'est sans doute sur cette base qu'ils se sont prononcés. Nous examinerons toutefois le bruit occasionnel de façon plus approfondie dans le prochain chapitre, et nous montrerons, en utilisant d'autres exemples, de quelle manière on peut le distinguer de la partie stable du bruit de pattern.

---

Pour résumer, nous avons identifié plusieurs types de bruit. Le *bruit systémique*, celui qu'on observe quand on considère un système où de nombreux juges rendent des jugements multiples, est la variabilité indésirable dans les jugements de la même affaire. Nous en avons identifié les deux grandes composantes, qui peuvent être séparées dès lors qu'on dispose des jugements rendus par les mêmes individus sur des cas multiples :

- Le *bruit de niveau* : c'est la variabilité du niveau moyen des jugements produits par des juges différents sur un ensemble de cas.
- Le *bruit de pattern* : c'est la variabilité dans les réactions des juges à des cas particuliers, indépendamment du niveau moyen de leurs jugements.

Dans l'étude que nous avons prise pour exemple, les volumes de bruit de niveau et de bruit de pattern étaient approximativement égaux. Par ailleurs, la composante que nous avons identifiée comme du bruit de pattern contient certainement un peu de *bruit occasionnel*, qui peut être assimilé à de l'erreur aléatoire.

---

L’audit de bruit du système judiciaire nous a servi d’illustration, mais la même analyse pourrait s’appliquer à tout audit de bruit, dans l’entreprise, en médecine, dans le système éducatif, dans un gouvernement ou ailleurs. Le bruit de niveau et le bruit de pattern (avec sa part de bruit occasionnel) contribuent tous les deux au bruit systémique. Nous allons les retrouver à de nombreuses reprises dans ce livre.

### À propos de l’analyse du bruit

« Le bruit de niveau, c’est le fait que les juges aient des niveaux de sévérité différents. »
« Le bruit de pattern, c’est le fait qu’ils ne soient pas d’accord entre eux sur le point de savoir quels prévenus méritent des peines plus sévères. »
« Et une partie du bruit de pattern, c’est du bruit occasionnel, quand les juges ne sont pas d’accord avec eux-mêmes. »
« Dans un monde parfait, les prévenus auraient affaire à la justice ; dans le monde actuel, ils ont affaire à un système judiciaire sujet au bruit. »

CHAPITRE 7

# Le bruit occasionnel

Un basketteur professionnel se prépare pour un lancer franc. Il est sur la ligne de lancer. Il se concentre. Il tire. Cette séquence de mouvements qu'il effectue, il l'a répétée un nombre incalculable de fois, à l'entraînement comme en match. Va-t-il marquer ?

Nous ne le savons pas et lui non plus. Dans les matches de la NBA (le championnat professionnel américain), les joueurs réussissent environ trois lancers francs sur quatre. Certains joueurs, évidemment, sont meilleurs que les autres, mais aucun n'atteint 100 % de réussite. Les meilleurs de tous les temps réussissent un peu plus de 90 % de leurs lancers francs. (Au moment où sont écrites ces lignes, ce sont Stephen « Steph » Curry, Steve Nash et Mark Price[1].) Les moins bons sont aux alentours de 50 % (le grand Shaquille O'Neal[2], par exemple, ne marquait que 53 % de ses lancers francs). Bien que le panier soit toujours exactement à 3,05 mètres du sol et à 4,60 mètres de la ligne de lancer franc, et que le ballon pèse toujours 624 grammes, répéter la suite précise des gestes nécessaires pour marquer n'est pas chose aisée. On peut s'attendre à une forte variabilité, non seulement entre les joueurs, mais aussi chez chacun d'entre eux. Le lancer franc est une forme de loterie. Certes, la probabilité de gagner est bien plus grande quand le tireur s'appelle Curry plutôt qu'O'Neal, mais c'est une loterie tout de même.

D’où vient cette variabilité ? D’innombrables facteurs peuvent influencer le joueur sur la ligne de lancer franc : la fatigue d’une longue partie, la pression d’un score serré, les hourras de ses fans ou les huées de ceux de l’équipe adverse. Mais, en vérité, il y a peu de chances que nous sachions jamais faire la part de ces différents facteurs. La variabilité de la performance d’un basketteur est une forme de bruit.

## *Une seconde loterie*

La variabilité des lancers francs ou d’autres processus physiques n’a rien d’étonnant. Nous sommes habitués à la variabilité dans nos corps : notre rythme cardiaque, notre pression artérielle, nos réflexes, le ton de notre voix, le tremblement de nos mains sont variables d’un moment à l’autre. Et quels que soient nos efforts pour produire la même signature, elle est toujours légèrement différente à chaque fois que nous signons un chèque.

Il est moins facile d’observer la variabilité de nos esprits. Bien sûr, tout le monde sait qu’on peut changer d’avis, même en l’absence d’informations nouvelles. Le film qui nous a fait rire aux éclats la veille nous paraît aujourd’hui médiocre. La personne que nous jugions sévèrement ce matin semble mériter ce soir notre indulgence. Un argument que nous n’avions pas apprécié ou compris nous paraît désormais essentiel. Mais, comme le suggèrent ces exemples, ces changements sont généralement associés à des questions relativement mineures et largement subjectives.

En réalité, nos opinions changent souvent sans raison apparente, même quand il s’agit des jugements réfléchis et prudents que produisent des experts. Il est ainsi fréquent que des médecins formulent des diagnostics très différents quand le même cas leur est présenté deux fois (voir le chapitre 22). Dans un grand concours de vins aux États-Unis, quand les mêmes œnologues goûtèrent les mêmes vins à deux reprises, 18 %

seulement des vins reçurent une note identique (généralement, les plus mauvais[3]). Un spécialiste de la police scientifique peut arriver à des conclusions différentes quand il examine les mêmes empreintes digitales à quelques semaines d'intervalle (voir le chapitre 20). Des informaticiens expérimentés peuvent proposer des estimations très différentes du temps nécessaire à l'accomplissement d'une tâche identique quand on leur soumet le même problème en deux occasions différentes[4]. En somme, de même qu'un basketteur ne lance jamais deux fois le ballon exactement de la même façon, nous ne produisons pas toujours des jugements identiques quand nous sommes confrontés deux fois aux mêmes données.

Nous avons parlé de loterie pour décrire le processus par lequel sont choisis un assureur, un juge ou un médecin. Cette loterie, en désignant un professionnel, produit du bruit systémique. Le bruit occasionnel est le résultat d'une seconde loterie, qui détermine le moment où ce professionnel produit un jugement. Avec ce moment viennent son humeur, la série de cas qu'il vient de traiter et qu'il a encore à l'esprit, et d'innombrables autres facteurs.

Plus encore que la première, cette seconde loterie demeure généralement invisible. Dans la compagnie d'assurances, par exemple, nous pouvons imaginer assez facilement que si la première loterie avait sélectionné un autre souscripteur, la prime fixée aurait été différente. Mais pour imaginer les jugements différents que le *même* souscripteur aurait pu produire, nous devons nous engager dans un raisonnement contrefactuel et abstrait. Ce que nous devons nous représenter, c'est que le jugement produit a été choisi au sein d'une constellation de possibilités. Le bruit occasionnel est la variabilité entre ces possibilités restées invisibles.

## *Mesurer le bruit occasionnel*

Les mêmes raisons qui font que l'existence du bruit occasionnel nous étonne en rendent la mesure difficile. Quand des gens se forment une opinion professionnelle en prenant toutes les précautions possibles, ils y associent les raisons qui justifient leur point de vue. Pressés d'expliquer leur jugement, ils le défendront avec des arguments convaincants. Et s'ils reconnaissent un problème qui leur a déjà été présenté, ils vont bien sûr reproduire leur première réponse, non seulement pour s'éviter un effort inutile, mais aussi pour faire preuve de cohérence. Si, par exemple, un professeur relit la dissertation sur laquelle il a inscrit, une semaine plus tôt, une excellente note, il y a peu de chances qu'il lui donne une note radicalement différente.

Cette tendance à la cohérence explique qu'il soit difficile d'obtenir des mesures directes du bruit occasionnel dans les domaines où les cas sont facilement mémorisables. Si vous montrez à un souscripteur d'assurances ou à un juge pénal un cas qu'il a récemment étudié, il le reconnaîtra probablement, et se contentera de répéter son jugement initial. Une synthèse des recherches faites sur la variabilité intra-individuelle dans les jugements professionnels (également appelée *stabilité* ou *fidélité test-retest*) a ainsi recensé de nombreuses études dans lesquelles des experts étaient appelés à produire le même jugement deux fois au cours de la même session de travail. Vous ne serez pas étonné d'apprendre qu'ils étaient généralement d'accord avec eux-mêmes[5].

Pour contourner ce problème, il faut avoir recours à des stimuli que les experts ne peuvent pas reconnaître. C'était le cas dans les exemples de bruit occasionnel que nous avons mentionnés plus haut. Les experts en empreintes devaient se prononcer sur des spécimens qu'ils avaient déjà analysés et les informaticiens sur des tâches qu'ils avaient déjà étudiées – mais à plusieurs semaines ou plusieurs mois d'intervalle. Quant aux œnologues, ils dégustaient les vins à l'aveugle. Et dans tous les cas, ces

professionnels ne savaient pas qu'on leur soumettait des problèmes qu'ils avaient déjà rencontrés.

Il existe un autre moyen, indirect, de mesurer le bruit occasionnel : les méthodes économétriques appuyées sur de très grandes masses de données (« big data »). En utilisant un grand échantillon de décisions professionnelles, il est parfois possible de quantifier l'impact qu'ont eu sur ces décisions des facteurs occasionnels non pertinents, comme l'heure de la journée ou la température extérieure. Si l'on découvre que ces facteurs non pertinents ont des effets statistiquement significatifs, on est en présence de bruit occasionnel. Bien sûr, on ne pourra jamais inventorier de cette manière toutes les sources exogènes de bruit occasionnel, mais celles que l'on peut détecter nous donnent déjà une indication de la très grande variété de ces sources de bruit. Et pour minimiser le bruit occasionnel, il est fort utile d'identifier certains des facteurs qui le produisent.

## *La foule intérieure*

À votre avis, quelle proportion des aéroports de la planète se trouve aux États-Unis ? Une réponse vous vient probablement déjà à l'esprit. Mais elle ne vous est pas venue de la même façon que vous vous rappelleriez votre âge ou votre numéro de téléphone. Vous savez que le chiffre que vous avez proposé n'est qu'une estimation. Vous n'avez pas choisi ce chiffre au hasard : « 1 % » ou « 99 % » seraient clairement de mauvaises réponses. Mais le chiffre qui vous est venu à l'esprit n'est qu'une possibilité parmi un grand nombre d'autres que vous ne considéreriez pas comme ridicules. Si quelqu'un ajoutait ou soustrayait un point de pourcentage à votre réponse, vous ne jugeriez sans doute pas sa réponse beaucoup moins plausible que la vôtre. (Pour les curieux, la réponse exacte est 32 %[6].)

Deux chercheurs, Edward Vul et Harold Pashler, ont posé cette question (et beaucoup d'autres) à un échantillon de participants. Après avoir répondu, les sujets avaient la surprise de découvrir qu'on leur demandait de répondre une seconde fois… à la même question[7] ! L'hypothèse que Vul et Pashler voulaient tester peut sembler étrange : ils supposaient que la *moyenne* des deux réponses successives serait plus juste que chacune des deux.

Les données qu'ils recueillirent leur donnèrent raison. En général, la première réponse était plus proche de la vérité que la seconde, mais la meilleure estimation était bien la moyenne des deux.

L'idée de Vul et Pashler était inspirée par un phénomène bien connu, appelé *sagesse des foules* : faire la moyenne de jugements produits de manière indépendante par des personnes différentes produit généralement un jugement plus exact. En 1907, Francis Galton, cousin de Darwin et savant célèbre, demanda à 787 villageois présents à une foire agricole d'estimer le poids d'un bœuf qui avait été primé. Aucun des participants ne trouva précisément le poids du bœuf, qui était de 1 198 livres, mais la moyenne de leurs estimations était de 1 200, soit 2 livres de plus seulement, et la médiane (1 207) était elle aussi très proche. Les villageois formaient une « foule sage » au sens où, si leurs estimations individuelles étaient sujettes au bruit, elles n'étaient pas biaisées : l'erreur moyenne était quasi nulle. Galton, qui n'avait que peu de respect pour les gens ordinaires, s'avoua lui-même surpris de ces résultats. Comme il le reconnut à contrecœur, « un jugement démocratique » s'avérait « plus fiable qu'on n'aurait pu le penser ».

On retrouve l'effet « sagesse des foules » dans des centaines de situations. Bien sûr, quand les questions sont si difficiles que seuls des experts peuvent approcher de la bonne réponse, les foules ne tombent pas nécessairement très juste. Mais si l'on demande à un grand nombre de gens de deviner le nombre de bonbons dans un bocal transparent, de prévoir la

température d'ici une semaine ou d'estimer la distance entre deux villes de leur pays, leur réponse médiane est généralement assez proche de la vérité[8]. Le mécanisme statistique à l'œuvre est élémentaire : en calculant la moyenne de jugements (ou de mesures) indépendants, on produit un nouveau jugement, qui est nécessairement moins sujet au bruit que les jugements individuels, même s'il n'est pas moins biaisé[9].

L'expérience de Vul et Pashler visait à déterminer si le même effet permet aussi de réduire le bruit occasionnel. Puisqu'on se rapproche de la vérité en combinant les réponses de plusieurs personnes, peut-on en faire autant en combinant deux réponses successives d'une même personne ? La réponse, nous l'avons vu, est positive. Vul et Pashler ont donné à cet effet un nom évocateur : la *foule intérieure*.

Certes, on n'améliore pas autant les jugements en prenant la moyenne des deux réponses d'une même personne qu'en utilisant une seconde opinion indépendante. Comme l'écrivent Vul et Pashler, « le gain obtenu en vous posant la même question deux fois est environ un dixième de celui que vous obtiendriez en demandant une seconde opinion à quelqu'un d'autre ». L'amélioration est modeste. Mais on peut obtenir un effet plus important en patientant un peu avant de reposer la question. Quand Vul et Pashler ont attendu trois semaines avant de reposer la question à leurs sujets, le gain est passé à un tiers environ de ce qu'apporterait une seconde opinion indépendante. Ce n'est pas rien, si l'on considère que la technique ne nécessite aucune information supplémentaire ni aucune aide extérieure. À tout le moins, ce résultat donne une certaine légitimité à un conseil vieux comme le monde : « Repensons-y demain ; la nuit porte conseil. »

Indépendamment des travaux de Vul et Pahsler et quasi simultanément, deux chercheurs allemands, Stefan Herzog et Ralph Hertwig, ont proposé une manière différente de tirer profit du même principe[10]. Au lieu de demander simplement à leurs sujets de produire librement une seconde estimation, ils les ont encouragés à en produire une qui soit aussi différente

que possible de la première, tout en demeurant plausible. Il fallait pour cela que les sujets réfléchissent activement aux informations qu'ils n'avaient pas prises en considération la première fois, et qui pourraient les amener à reconsidérer leur point de vue. Voici les instructions qui leur étaient données :

*1) Partez du principe que votre première estimation est loin du compte.*
*2) Demandez-vous pourquoi. Lesquelles de vos hypothèses et de vos considérations peuvent avoir été fausses ?*
*3) Qu'est-ce qu'impliquent ces nouvelles considérations ? Votre première estimation était-elle trop élevée ou trop basse ?*
*4) À partir de cette perspective nouvelle, proposez une nouvelle estimation.*

Comme Vul et Pahsler, Herzog et Hertwig firent ensuite la moyenne des deux estimations ainsi produites. Leur technique (nommée *réestimation dialectique*) produisit des améliorations plus importantes qu'une simple demande de seconde estimation suivant immédiatement la première. Comme les participants s'obligeaient à examiner la question sous un nouveau jour, ils mobilisaient, si l'on peut dire, une version d'eux-mêmes plus différente de la première : deux « membres » de leur « foule intérieure » plus éloignés l'un de l'autre. Par conséquent, leur moyenne produisait une estimation plus proche de la vérité. Le gain d'exactitude obtenu en prenant la moyenne de deux estimations « dialectiques » immédiatement consécutives était d'environ la moitié de celui qu'apporte le jugement d'une personne différente.

Herzog et Hertwig résument simplement le conseil méthodologique qu'on peut donner, sur la base de ces résultats, à toute personne en situation de décision. Avant tout, si vous pouvez obtenir l'opinion indépendante de plusieurs autres personnes, faites-le : la sagesse de cette vraie « foule »

améliorera très probablement votre jugement. Si ce n'est pas possible, produisez vous-même un nouveau jugement pour créer une « foule intérieure ». Vous pouvez le faire après qu'un peu de temps a passé, ce qui vous permet de prendre un peu de distance avec votre premier jugement. Vous pouvez également vous efforcer de penser contre vous-même et de trouver une autre façon de voir le problème (la réestimation dialectique). Enfin et surtout, qu'il s'agisse d'une foule réelle ou d'une foule intérieure, et à moins que vous n'ayez de très bonnes raisons d'accorder plus de poids à l'une des estimations plutôt qu'à l'autre, vous aurez tout intérêt à en faire la moyenne.

Au-delà du conseil pratique, ces travaux confirment une idée fondamentale sur le jugement : comme l'écrivent Vul et Pashler, « les réponses faites par un sujet ne sont pas sélectionnées de façon déterministe sur la base de toutes les connaissances dont il dispose, mais choisies aléatoirement dans l'éventail des possibilités qu'il a mentalement générées[11] ». Vous retrouvez dans cette observation l'expérience que vous avez faite quand vous avez répondu à la question sur les aéroports aux États-Unis : votre première estimation n'a pas mobilisé l'ensemble de vos connaissances, ni même les plus pertinentes. Ce n'était qu'une réponse parmi la constellation de réponses possibles que votre esprit aurait pu produire. Plus généralement, la variabilité que nous observons dans les jugements du même problème par la même personne n'est pas une bizarrerie suscitée par l'analyse de problèmes très complexes : chacun de nos jugements émane d'un individu choisi dans notre foule intérieure, et aurait tout aussi bien pu être produit par un autre de ses membres. Le bruit occasionnel affecte tous nos jugements, tout le temps.

## *Les sources du bruit occasionnel*

Il y a au moins une source de bruit occasionnel que nous avons tous remarquée : l'humeur. Il nous est tous arrivé un jour que notre jugement soit affecté par l'humeur dans laquelle nous nous sentions ; et nous savons bien que celui des autres varie aussi en fonction de leur humeur.

L'effet des humeurs sur le jugement a fait l'objet de très nombreuses recherches en psychologie. Il est étonnamment facile de rendre les gens temporairement gais ou tristes, et de mesurer la variabilité de leurs décisions et de leurs jugements après les avoir ainsi influencés. On peut par exemple demander à des sujets de rédiger un paragraphe évoquant un souvenir heureux ou malheureux ; ou leur faire visionner un court extrait d'une comédie ou d'un mélo.

Parmi les psychologues qui ont étudié les effets de la manipulation de l'humeur, le plus prolifique est sans doute l'Australien Joseph Forgas[12], qui a publié près d'une centaine d'articles scientifiques sur le sujet. Certains travaux de Forgas confirment ce que vous savez déjà : les gens qui sont de bonne humeur sont généralement plus positifs. Ils évoquent plus facilement des souvenirs heureux que des souvenirs malheureux ; sont plus positifs sur le comportement d'autrui ; sont plus généreux et plus serviables, etc. L'humeur négative a les effets inverses. Comme l'écrit Forgas : « Le même sourire qui est considéré comme amical par quelqu'un de bonne humeur peut être jugé importun par quelqu'un de mauvaise humeur ; et discuter du temps qu'il fait semble plaisant à une personne de bonne humeur mais irritant quand on est de mauvaise humeur[13]. »

Autrement dit, l'humeur a une influence mesurable sur *ce que* nous pensons : ce que nous remarquons dans notre environnement, les souvenirs qui nous reviennent, et la manière dont nous interprétons ces signaux. Mais l'humeur a un autre effet, plus surprenant : elle change aussi notre *façon* de penser. Et là, les effets ne sont pas ceux que vous croyez. La bonne humeur n'a pas que des avantages, et la mauvaise humeur n'a pas que des

inconvénients. Leurs coûts et leurs bénéfices sont propres à chaque situation.

Dans une situation de négociation, par exemple, la bonne humeur est un atout. Les gens de bonne humeur sont plus coopératifs et suscitent la réciprocité. Ils ont tendance à obtenir de meilleurs résultats que les négociateurs grincheux. (Bien sûr, les négociations réussies rendent aussi les gens heureux, mais dans les expériences dont il est ici question, l'humeur n'est pas provoquée par ce qui se passe pendant la négociation, puisqu'elle est induite au préalable.) Une précision tout de même : les négociateurs qui perdent leur bonne humeur et se mettent en colère en cours de négociation ont souvent de bons résultats. À retenir pour le jour où vous ferez face à un adversaire particulièrement têtu [14] !

Dans d'autres situations, en revanche, la bonne humeur est un handicap. En particulier, quand nous sommes de bonne humeur, nous sommes plus enclins à nous fier à nos premières impressions sans les remettre en cause. Dans l'une des études de Forgas, les participants devaient lire un court texte de philosophie auquel était accolée une photographie de son auteur supposé [15]. Pour certains participants, la photo représentait le stéréotype du professeur de philo : un homme d'âge mûr portant des lunettes. Pour d'autres, la photo était celle d'une jeune femme. Comme vous le devinez, l'expérience visait à mesurer la sensibilité des lecteurs aux stéréotypes : notent-ils plus favorablement le texte quand il est attribué à un homme mûr que quand il est signé d'une jeune femme ? Hélas, la réponse est oui, bien sûr. Mais surtout, la différence entre les notes est plus grande quand les sujets, avant de lire le texte, sont de bonne humeur ! Les gens de bonne humeur sont davantage victimes de leurs biais.

D'autres études ont testé l'effet de l'humeur sur la crédulité. Gordon Pennycook et ses collègues ont réalisé de nombreuses études sur les réactions à des affirmations absurdes, générées par l'assemblage aléatoire de verbes, d'épithètes et de substantifs piochés dans les propos de gourous

populaires. Les phrases produites par ce procédé sont grammaticalement correctes mais dénuées de sens, comme : « La plénitude apaise les phénomènes infinis » ou « Le sens caché transforme la beauté abstraite[16] ». La propension à trouver du sens à ce genre de phrases creuses est une tendance connue, appelée la *réceptivité au bullshit*. (Le *bullshit*, forme particulière de « connerie » ou d'« enfumage », est devenu une sorte de terme technique depuis qu'Harry Frankfurt, un philosophe de Princeton, lui a consacré un livre où il le distingue d'autres types de représentations erronées[17].)

Assurément, certaines personnes sont plus réceptives que d'autres au *bullshit*. Elles peuvent être impressionnées par « des affirmations apparemment puissantes présentées comme profondes et vraies, mais qui sont, en réalité, totalement creuses[18] ». Là encore, cependant, cette crédulité n'est pas seulement fonction de dispositions permanentes et immuables : susciter la bonne humeur rend les gens plus réceptifs au *bullshit.* La bonne humeur les rend d'ailleurs plus crédules en général : ils sont moins aptes à détecter les efforts pour les tromper ou les informations de nature à les induire en erreur[19]. Inversement, les témoins oculaires qui sont exposés à des informations erronées sont plus à même de ne pas en tenir compte – et d'éviter ainsi un faux témoignage – quand ils sont de mauvaise humeur[20].

Même les jugements moraux sont fortement influencés par l'humeur. Des chercheurs ont par exemple soumis à leurs sujets un problème classique de philosophie morale : le dilemme du tramway, dans sa variante dite « de la passerelle » ou « de l'homme obèse[21] ». L'expérience de pensée est la suivante. Un tramway devenu incontrôlable menace, s'il poursuit sa course, de tuer cinq personnes qui se trouvent sur sa voie. Vous contemplez la scène depuis une passerelle en surplomb de la voie, où se trouve également un homme très corpulent. Le seul moyen d'intervenir est de pousser cet homme du haut de la passerelle pour qu'il tombe sur la voie et fasse

dérailler le tramway. Si vous décidez de pousser l'homme, il mourra, mais les cinq autres personnes seront épargnées.

Le dilemme du tramway illustre la tension entre deux doctrines éthiques opposées. Le calcul utilitariste, associé au philosophe anglais Jeremy Bentham, suggère d'intervenir : la perte d'une vie est préférable à la perte de cinq vies. Mais l'éthique déontologique, associée au philosophe allemand Emmanuel Kant, interdit de tuer un être humain, même pour en sauver plusieurs autres. Alors que ce dilemme est généralement présenté comme le choix d'actionner ou non un aiguillage, sa variante de la passerelle a pour particularité de présenter une forte composante émotionnelle : pousser un homme d'un pont pour qu'un tramway lui roule dessus est un acte particulièrement répugnant. Pour choisir ici d'intervenir, il faut donc non seulement adopter le calcul utilitariste, mais aussi surmonter son aversion pour un acte physiquement violent contre un inconnu. Seule une petite minorité de participants (dans cette étude, moins d'un sur dix) se disent généralement prêts à le faire.

Chose étonnante : ce choix moral est, lui aussi, très sensible à l'humeur. Quand les sujets ont visionné une vidéo de cinq minutes choisie pour les mettre dans une humeur positive, ils sont trois fois plus nombreux à déclarer qu'ils pousseraient l'homme du pont. Considérez-vous le commandement « Tu ne tueras point » comme un principe absolu, ou êtes-vous au contraire prêt à sacrifier un inconnu pour en sauver cinq autres ? La réponse devrait, en principe, refléter vos valeurs les plus profondes. Il semble pourtant qu'elle dépende de la vidéo que vous venez de regarder.

Si nous décrivons en détail ces études sur l'effet de l'humeur, c'est pour illustrer une idée qui nous paraît capitale : *vous n'êtes pas la même personne tout le temps*. Quand votre humeur varie (ce dont vous avez bien sûr conscience), le fonctionnement de votre appareil cognitif varie aussi (et cela, vous n'en avez pas forcément conscience). Lorsque vous êtes confronté à un problème de jugement complexe, il est tout à fait possible

que votre humeur du moment influence votre manière d'aborder le problème et les conclusions auxquelles vous allez arriver, même si vous êtes convaincu que votre humeur ne peut en aucun cas avoir ce pouvoir, et même si vous pouvez sans difficulté trouver des justifications à votre jugement. En un mot, vous êtes sujet au bruit occasionnel.

Beaucoup d'autres facteurs exogènes ne devraient pas perturber les jugements professionnels, mais les influencent quand même. Deux de ces sources de bruit occasionnel sont bien connues : le stress et la fatigue. Une étude portant sur près de 700 000 consultations chez des médecins généralistes, par exemple, a montré qu'ils prescrivaient beaucoup plus d'opioïdes à la fin d'une longue journée[22]. Bien sûr, il n'y a aucune raison que les patients dont le rendez-vous est à 16 heures souffrent davantage que ceux reçus à 9 heures. Et le fait que le médecin soit en retard dans son planning ne devrait pas influencer non plus ses choix thérapeutiques. D'ailleurs, les prescriptions d'autres traitements de la douleur, comme les séances de kinésithérapie ou les anti-inflammatoires non stéroïdiens, ne montrent pas la même sensibilité à l'horaire. Tout se passe comme si les médecins, sous la pression du retard et de la fatigue accumulés au long de la journée, étaient plus enclins à choisir des solutions « rapides », en dépit de leurs sérieux inconvénients. De la même manière, d'autres études ont montré qu'en fin de journée, les médecins ont davantage tendance à prescrire des antibiotiques[23] et sont moins nombreux à prescrire des vaccins antigrippaux[24].

Même le temps qu'il fait exerce une influence mesurable sur les jugements professionnels. Comme ces jugements sont souvent rendus dans des salles équipées d'air conditionné, l'effet de la météo est probablement indirect et « médié » par l'humeur : la météo n'affecte pas directement les décisions, mais modifie l'humeur de la personne qui décide, ce qui change sa manière de décider. Le mauvais temps est associé à une meilleure mémoire[25], les peines judiciaires ont tendance à être plus sévères quand il

fait chaud dehors, et les marchés financiers sont affectés par le beau temps[26]. Dans certains cas, l'effet de la météo est encore plus étrange. Uri Simonsohn a montré que les responsables des admissions à l'université prêtent davantage attention au cursus scolaire des candidats quand le temps est nuageux, tandis que, quand le temps est ensoleillé, ils sont plus sensibles aux activités extrascolaires. Ce que résume de manière lapidaire le titre de l'article : « Les polards brillent dans le brouillard[27]. »

Il existe une autre source importante de variabilité du jugement : les effets de séquence, c'est-à-dire l'ordre dans lequel les cas se présentent. Quand un individu examine une affaire, les décisions qui l'ont immédiatement précédée lui servent implicitement de cadre de référence. Beaucoup de professionnels prennent chaque jour une série de décisions successives : c'est le cas des juges qui voient défiler les prévenus, mais aussi des banquiers qui examinent des demandes de crédit ou des arbitres de base-ball qui jugent une succession de balles. Tous semblent enclins à restaurer une forme d'équilibre : après une série de décisions allant dans le même sens, ils ont plus tendance à décider dans le sens opposé que ne le justifierait l'examen de chaque décision prise isolément. Les erreurs (et l'injustice) sont, de ce fait, inévitables. Aux États-Unis, les juges des demandes d'asile, par exemple, ont 19 % moins de chances d'accorder l'asile à un requérant quand les deux demandes précédentes ont été satisfaites[28]. Et la personne qui sollicite un prêt aura plus de chances de se le voir accorder si les deux demandes précédentes ont été refusées, et plus de chances d'essuyer un refus si les deux demandes précédentes ont été satisfaites.

Ce comportement témoigne d'un biais cognitif appelé l'*erreur du parieur* : dans toute suite aléatoire, par exemple une suite de résultats de tirages à pile ou face, il se produit inévitablement des « séries », comme quand on tire « pile » trois, quatre, ou cinq fois de suite. Ces séries ont une fréquence naturelle d'occurrence : le hasard produit spontanément des

séries. Mais nous avons tendance à sous-estimer la fréquence à laquelle elles apparaissent. Les séries nous semblent « anormales », et nous avons donc tendance à vouloir les corriger.

## *Quantifier le bruit occasionnel*

Quelle est la part du bruit occasionnel dans le bruit systémique total ? Si aucun chiffre unique ne peut bien sûr s'appliquer à toutes les situations, une règle générale se dégage : quantitativement, les effets décrits dans ce chapitre sont inférieurs aux différences stables entre individus, que ce soit dans le niveau moyen de leurs jugements ou dans leurs patterns personnels.

Nous venons de mentionner, par exemple, que les chances pour un demandeur d'asile d'être admis aux États-Unis diminuent de 19 % si le juge qui examine sa demande vient d'accorder l'asile à deux autres requérants. Cette variabilité est certainement troublante. Mais ce n'est rien en comparaison de la variabilité *entre* juges : dans le même tribunal de Miami, Jaya Ramji-Nogales et ses coauteurs ont observé que l'un des juges accordait l'asile à 88 % des demandeurs et un autre à seulement 5 %. (Il ne s'agit pas d'un audit de bruit mais de données réelles, ce qui veut dire que les demandeurs étaient tous différents ; cependant, ils étaient déférés devant tel ou tel juge de façon quasi aléatoire, et les auteurs ont vérifié que les écarts ne s'expliquaient pas par des différences de pays d'origine.) Compte tenu de pareilles disparités, le fait que l'un de ces deux chiffres puisse être réduit d'un cinquième environ par les circonstances du jugement semble presque insignifiant[29].

De même, les examinateurs d'empreintes digitales et les médecins sont quelquefois en désaccord avec eux-mêmes, mais moins souvent qu'ils ne le sont avec d'autres. Dans tous les cas que nous avons examinés et où la part du bruit occasionnel dans le bruit systémique total pouvait être mesurée, la

contribution du bruit occasionnel était plus faible que celle des différences entre individus.

Ou pour le dire autrement : vous n'êtes pas toujours la même personne, et vous êtes moins cohérent dans le temps, moins semblable à vous-même, que vous ne l'imaginez. Mais, si cette variabilité vous inquiète, un autre constat pourra peut-être vous rassurer : vous ressemblez quand même davantage à la personne que vous étiez hier qu'à une autre personne aujourd'hui.

## *Aux racines du bruit occasionnel*

Humeur, fatigue, stress, météo, effets de séquence : nombreux sont les facteurs susceptibles de déclencher des variations indésirables dans le jugement du même problème par le même individu. On pourrait imaginer de construire un environnement de décision dans lequel tous ces facteurs exogènes seraient connus et contrôlés. En théorie au moins, un tel dispositif devrait réduire le bruit occasionnel. Mais même cela ne suffirait pas à l'éliminer totalement.

Pour le comprendre, il faut évoquer les travaux de Michael Kahana et de ses collègues de l'Université de Pennsylvanie sur la performance en matière de mémoire. (La mémorisation ne relève certes pas du jugement tel que nous l'avons défini, mais c'est une tâche cognitive qui se prête particulièrement bien au contrôle des conditions d'expérience et à la mesure des variations de performance.) Dans une de leurs études[30], ils ont demandé à soixante-dix-neuf sujets de participer à une analyse remarquablement détaillée de leurs performances mnésiques. Les sujets ont pris part, jour après jour, à vingt-trois sessions successives, pendant lesquelles ils ont dû mémoriser des mots issus de vingt-quatre listes différentes comprenant

vingt-quatre mots chacune. La performance mnésique était définie par le pourcentage de mots mémorisés.

Kahana et ses collègues ne se sont pas intéressés à la différence entre les sujets (la variation interindividuelle), mais plutôt aux indicateurs de variabilité intra-individuelle : la fluctuation dans les performances de chacun. Ces performances seraient-elles influencées par la forme dans laquelle les sujets se sentaient ? Par la qualité de leur sommeil la nuit qui avait précédé ? Par l'heure de la journée à laquelle la session avait lieu ? S'amélioreraient-elles avec la pratique, d'une session à la suivante ? Se dégraderaient-elles entre le début et la fin de chaque session, à cause de la fatigue ou de l'ennui ? Certaines listes de mots seraient-elles plus faciles à mémoriser que d'autres ?

À toutes ces questions, la réponse était, pour faire simple, la même : « oui… mais pas beaucoup ». Un modèle intégrant toutes ces variables n'expliquait que 11 % de la variation des performances d'un sujet donné. Comme le soulignent les chercheurs, « nous avons été stupéfaits de voir qu'une très grande part de la variabilité demeurait inexpliquée, même une fois que les effets de toutes nos variables prédictives étaient pris en compte ». En d'autres termes, même dans un environnement étroitement contrôlé, l'origine du bruit occasionnel restait pour l'essentiel mystérieuse.

De toutes les variables étudiées par les chercheurs, celle qui prédisait le mieux la performance d'un sujet dans la mémorisation d'une liste donnée n'était pas un facteur exogène. Le meilleur prédicteur du score d'un sujet sur une liste de mots donnée était son score sur la liste qui l'avait immédiatement précédée. Une liste « réussie » avait tendance à être suivie par une autre relativement réussie, et une liste médiocre par une autre liste médiocre. La performance ne variait pas de façon aléatoire d'une liste à l'autre : durant chacune des sessions, elle fluctuait avec le temps, sans cause extérieure évidente.

Ces résultats suggèrent que la performance mnésique tient en grande partie, écrivent Kahana et ses coauteurs, à « l'efficacité des processus neuronaux endogènes qui régissent la fonction mémorielle ». Autrement dit, la variabilité de l'efficacité de notre esprit d'un moment à l'autre n'est pas seulement due à des facteurs exogènes, comme le temps qu'il fait ou une intervention perturbatrice. C'est une caractéristique du fonctionnement de notre cerveau.

Il est très probable que la variabilité intrinsèque dans le fonctionnement du cerveau affecte aussi la qualité de nos jugements, d'une façon que nous ne pouvons pas espérer contrôler. Cette variabilité implique que le bruit occasionnel ne peut guère être complètement éliminé. L'analogie avec le basketteur sur la ligne de lancer franc n'était donc pas aussi simpliste qu'elle a pu le sembler de prime abord. De même que les muscles du joueur n'exécutent jamais exactement le même geste, nos neurones n'opèrent jamais exactement de la même façon. Notre esprit est un instrument de mesure, mais il ne sera jamais parfait.

---

Le bruit occasionnel qui affecte nos jugements provient d'une multitude de facteurs que nous ne pouvons pas tous contrôler. Mais nous pouvons au moins en contrôler certains. Cet objectif prend une importance particulière quand les jugements sont produits en groupe, comme nous allons le voir dans le chapitre suivant.

## À propos du bruit occasionnel

« Le jugement est pareil au lancer franc : on a beau essayer de le répéter, il n'est jamais tout à fait identique. »
« Votre jugement dépend de l'humeur dans laquelle vous êtes, des cas dont vous avez discuté juste avant, et même du temps qu'il fait. Vous n'êtes pas toujours la même personne. La loterie qui choisit le moment du jugement joue un rôle important. »
« Vous avez fait un jugement ; vous auriez pu en faire un autre. Le jugement que vous avez exprimé est celui d'un individu choisi dans votre foule intérieure. »
« Si vous faisiez la moyenne de deux individus dans cette foule – c'est-à-dire de deux de vos jugements à des moments différents – le résultat serait meilleur. »
« Même s'il se peut que vous ne soyez pas la même personne que celle que vous étiez la semaine dernière, vous êtes moins différent de "vous" la semaine dernière que vous ne l'êtes de quelqu'un d'autre aujourd'hui. Le bruit occasionnel n'est pas la source de bruit systémique la plus importante. »

CHAPITRE 8

# Les groupes et le bruit

Le bruit dans le jugement individuel est déjà préoccupant, mais quand les jugements sont produits par des groupes, le problème du bruit prend une nouvelle dimension. Il est fréquent de voir des groupes confrontés à des problèmes similaires prendre des décisions très différentes, qu'il s'agisse de recrutement, de promotion, de restructuration d'entreprise, de réglementation environnementale, de sécurité nationale, de stratégie de communication, d'admission à l'université, ou de lancement de nouveaux produits.

Cette observation peut sembler paradoxale, puisque nous avons vu dans le précédent chapitre qu'agréger les jugements de multiples individus est une excellente manière de réduire le bruit. Mais la dynamique des groupes fait qu'ils peuvent tout aussi bien ajouter du bruit qu'en retrancher. Il existe, nous l'avons vu, des « foules sages », dont le jugement moyen est proche de la réponse correcte ; mais nous savons tous qu'il existe aussi des foules qui acclament des tyrans, qui alimentent des bulles spéculatives, qui croient à la magie ou qui sont sous l'emprise d'une illusion partagée.

En pratique, des différences mineures (qui parle le premier, qui s'exprime avec assurance, qui sourit ou fronce les sourcils à un moment donné) peuvent conduire un groupe à dire « oui » et un groupe quasi

identique à dire « non ». On peut le constater en observant le bruit qui sépare des groupes dont les jugements devraient être identiques ; et l'on verra qu'au sein d'un même groupe, le jugement final, comme celui des individus, n'est qu'une option choisie dans une constellation d'autres possibilités.

## *Du bruit dans la musique*

Commençons par un domaine qui, de prime abord, pourrait paraître éloigné du sujet : une étude à grande échelle, réalisée par le sociologue Matthew Salganik et ses coauteurs, sur les téléchargements de chansons[1]. En s'adossant à un site de téléchargement de taille moyenne, les expérimentateurs créèrent d'abord un groupe de contrôle de quelques milliers de personnes, qui pouvaient écouter et télécharger soixante-douze chansons de groupes nouveaux. Dotées de titres très imagés, les chansons étaient bien réelles : « Trapped inside an Orange Peel » (« Piégé par une pelure d'orange »), « Gnaw » (« Ronge »), ou encore « Pink Aggression » (« Agression rose »). Étrangement, certains de ces titres semblent d'ailleurs liés directement à nos préoccupations : « Best Mistakes » (« Meilleures erreurs »), « I Am Error » (« Je suis une erreur »), ou « The Belief Above the Answer » (« La Croyance au-dessus de la réponse »).

Dans ce groupe de contrôle, les utilisateurs n'avaient aucune information sur les préférences ou les téléchargements des autres utilisateurs. Ils devaient donc se forger leur propre opinion sur les chansons qu'ils aimaient avant de les télécharger. Mais, en parallèle, Salganik et ses collègues créèrent huit autres groupes, dans lesquels ils répartirent de manière aléatoire des milliers d'autres utilisateurs. La seule différence entre ces groupes et le groupe de contrôle était que les utilisateurs pouvaient y voir combien de membres du groupe avaient téléchargé chaque chanson. Si

« Best Mistakes » était très populaire dans un groupe, par exemple, les membres du groupe le savaient, et si personne ne le téléchargeait, ils le savaient aussi.

Comme, à ce détail près, la composition des groupes était quasiment identique, on peut considérer l'étude comme une analyse de huit mondes parallèles, ou comme une manière de rejouer huit fois l'histoire. À l'issue du processus, on pourrait donc s'attendre à ce que les résultats soient identiques : dans tous les groupes, les classements devraient être similaires, avec les bonnes chansons classées en premier et les mauvaises en dernier. En d'autres termes, il ne devrait pas y avoir de bruit entre les groupes. Mais Salganik et ses coauteurs se doutaient que le résultat serait bien différent, à cause d'un facteur de bruit dont ils voulaient mesurer l'effet : l'*influence sociale*.

De fait, la comparaison entre les groupes révéla des classements radicalement différents d'un groupe à un autre : une quantité phénoménale de bruit. Dans un groupe donné, « Best Mistakes » pouvait avoir un succès spectaculaire et « I Am Error » faire un flop ; dans un autre groupe, ce pouvait être exactement le contraire. Quand une chanson rencontrait le succès très tôt, elle finissait souvent très bien classée ; dans le cas contraire, le résultat pouvait être complètement différent.

La seule constante semblait concerner les extrêmes : les chansons les plus mauvaises (d'après le jugement du groupe de contrôle) ne finissaient jamais tout en haut du classement ; et les toutes meilleures ne se languissaient jamais tout en bas. Mais, à cette réserve près, à peu près n'importe quoi pouvait arriver. Comme l'écrivent les auteurs, « le niveau de popularité était bien plus imprévisible dans une situation où l'influence sociale joue un rôle que quand les jugements sont produits de manière indépendante ».

Autrement dit, l'influence sociale crée beaucoup de bruit entre les groupes. Cette observation en implique une autre : chaque groupe est lui

aussi sujet au bruit, en ce sens que son jugement sur une chanson donnée aurait pu être complètement différent si cette chanson avait été un peu plus populaire (ou un peu moins) au tout début de l'expérience.

Salganik et ses coauteurs ne s'arrêtèrent pas là. Dans une seconde série d'études, ils entreprirent de manipuler les jugements de groupe, en utilisant le caractère autoréalisateur de la popularité[2]. Dans l'une de ces expériences, ils n'hésitèrent pas à inverser purement et simplement le classement du groupe de contrôle, ce qui revenait à présenter aux participants un palmarès diamétralement opposé à la réalité : les chansons les moins appréciées étaient présentées comme étant les plus populaires, et *vice versa*. Que croyez-vous qu'il arriva ? La plupart des chansons les moins appréciées devinrent assez populaires ; tandis que la plupart des chansons qui, normalement, auraient rencontré le succès n'en eurent qu'assez peu. La popularité créait plus de popularité, et l'impopularité plus d'impopularité – même quand les chercheurs avaient menti sur la popularité initiale. Là aussi, la seule exception concernait un cas extrême : la chanson la plus appréciée par le groupe de contrôle (en d'autres termes, la meilleure), et qui était donc initialement présentée en queue de classement, parvint, avec le temps, à remonter son handicap.

Ces études sont riches d'enseignements sur les jugements des groupes en général. Imaginez un petit groupe – par exemple, une dizaine de personnes – qui discute d'un ambitieux nouveau projet. Si un ou deux partisans du projet sont les premiers à s'exprimer, il y a de bonnes chances qu'ils entraînent les autres dans leur sens. Et il en ira de même si ce sont deux sceptiques qui parlent en premier. Il suffit pour cela que les membres du groupe soient influencés les uns par les autres, ce qui est généralement le cas. C'est ainsi que des groupes en tout point comparables peuvent prendre des décisions très différentes.

## *Dans la musique… mais pas seulement*

Peut-être pensez-vous que nous allons trop loin en extrapolant l'exemple du téléchargement de chansons aux jugements d'autres groupes de toutes sortes. Mais des observations similaires ont été faites dans bien d'autres domaines[3]. La popularité des propositions de référendum au Royaume-Uni en offre un exemple frappant. Quand les citoyens décident d'apporter ou non leur appui à une proposition de référendum, ils devraient, en principe, s'attacher à évaluer les mérites de chaque proposition. Toutefois, on observe un phénomène similaire à celui mesuré dans les téléchargements de musique. Une flambée initiale de popularité se renforce d'elle-même, tandis qu'une proposition qui suscite peu d'intérêt le premier jour est quasiment condamnée. En politique comme en musique, l'influence sociale joue un rôle capital : selon que les individus connaissent ou non les jugements de leurs pairs, les résultats sont radicalement différents.

Inspirés par l'étude de Salganik, le sociologue Michael Macy[4] et ses collaborateurs se sont demandé si la même logique s'applique aux opinions politiques. Le fait que les opinions d'autres personnes sur une question donnée soient rendues visibles conditionne-t-il l'avis des électeurs démocrates ou républicains sur cette question ? En un mot, la réponse est oui. Quand des Démocrates, sur un groupe de discussion en ligne, voient que beaucoup de Démocrates soutiennent une position donnée, ils adoptent à leur tour cette position, ce qui finit par conduire la plupart des Démocrates du même groupe à l'approuver à leur tour. Mais si, dans un autre groupe de discussion en ligne, les Démocrates voient que la même position est populaire chez les Républicains, ils la rejettent, ce qui finit par amener la plupart des Démocrates du groupe à la rejeter. La même dynamique se retrouve chez les Républicains. En d'autres termes, il en va des positions politiques comme des goûts musicaux : le destin d'une opinion peut dépendre, bien plus qu'on ne l'imagine, de sa popularité initiale. Comme

l’écrivent les chercheurs, il suffit de « variations aléatoires chez un petit nombre de personnes qui se prononcent les premières » pour produire des effets considérables, qui peuvent faire basculer l’opinion d’un groupe important. Ce phénomène contribue également à ce que les électeurs d’un parti – qu’il s’agisse des Républicains ou des Démocrates – adoptent comme un seul homme un agrégat de positions qui n’ont en réalité rien de commun entre elles, sinon d’être majoritaires au sein de leur parti.

Autre exemple, qui concerne directement la décision de groupe en général : comment les commentaires postés en ligne sont-ils jugés par les internautes ? Avec ses collègues, Lev Muchnik, professeur à l’Université hébraïque de Jérusalem, a réalisé une expérience sur un site Internet qui publie des articles et qui permet aux internautes de publier des commentaires. Ces commentaires peuvent à leur tour faire l’objet de votes « pour » ou « contre[5] ». De façon automatique et artificielle, les chercheurs se sont arrangés pour que le premier vote donné à certains commentaires soit un vote « pour ». La manipulation ne concerne donc qu’un seul vote. Peut-être pensez-vous que l’importance de ce vote initial est vite diluée, et qu’après des centaines ou des milliers de visites et d’évaluations, il n’a plus beaucoup d’importance.

Si c’est le cas, vous vous trompez. Quand le premier vote est positif (et même quand il est totalement factice, comme c’est ici le cas), le deuxième vote a 32 % de chances de plus d’être à son tour positif. Mais ce qui est remarquable, c’est que cet effet persiste dans le temps. Après cinq mois, on peut encore mesurer l’effet de ce premier vote artificiel : il accroît de 25 % la note moyenne des commentaires. En somme, le premier vote fabrique du bruit : il a un effet disproportionné sur le score final. Notons au passage que, si l’étude de Muchnik porte sur des groupes de très grande taille, le même phénomène se produit dans des groupes plus petits, et même de façon encore plus spectaculaire, puisque le poids relatif du vote initial n’en est que plus grand.

Ces commentaires sur la dynamique des groupes nous ramènent à nos observations sur la sagesse des foules, et nous imposent de les nuancer. Souvenez-vous du principe : si vous prenez un groupe de grande taille et si vous posez une question à ses membres, il est très probable que la réponse moyenne sera proche de la réponse juste. Agréger les jugements peut être un excellent moyen de réduire à la fois le biais et l'erreur. Mais que se passe-t-il quand chaque membre du groupe est à l'écoute de l'avis des autres ? On pourrait penser que c'est un avantage. Après tout, nous avons tous beaucoup à apprendre les uns des autres ; et l'avis d'autrui peut nous éclairer quand nous sommes dans le doute. De fait – nous y reviendrons –, la délibération peut quelquefois contribuer à produire de bons jugements, quand les circonstances et la méthode de travail sont favorables à la mise en commun des connaissances. Mais il n'y a pas de sagesse des foules sans indépendance des jugements. Si les membres de la « foule » ne se prononcent pas indépendamment les uns des autres, et s'ils s'appuient au contraire sur l'opinion des autres pour se forger la leur, alors la foule manquera cruellement de sagesse.

Une autre étude démontre cette limite de la sagesse des foules[6]. Dans des tâches simples d'estimation – pour évaluer, par exemple, le nombre de crimes commis dans une ville, la croissance de la population sur une période donnée, ou la longueur d'une frontière –, la sagesse des foules fonctionne bien, du moins quand chaque individu fournit son estimation indépendamment des autres. Mais dès lors que les participants sont informés des estimations faites par leurs pairs – par exemple, quand on leur indique l'estimation moyenne faite par un sous-ensemble de douze personnes –, la sagesse des foules disparaît. Comme le notent les auteurs de l'étude, l'influence sociale est nuisible parce qu'elle réduit « la diversité du groupe sans diminuer l'erreur collective ». La sagesse des foules est puissante, mais fragile : il suffit d'un rien d'influence sociale pour qu'un effet « mouton de Panurge » la fasse disparaître.

## *Les effets de cascade*

Le phénomène impliqué dans les études que nous venons de décrire porte un nom : la *cascade informationnelle*. Ces cascades sont partout. Elles permettent de comprendre pourquoi – dans l'entreprise, en politique et ailleurs – des changements infimes ont parfois des effets très importants sur la décision des groupes.

Pour comprendre le fonctionnement d'une cascade informationnelle, imaginez un groupe de dix personnes réunies pour choisir entre trois candidats à un poste important : Thomas, Sam et Julie. Les participants à la réunion s'expriment, comme c'est l'usage, l'un après l'autre. Chacun, naturellement, est à l'écoute du jugement de ses pairs.

Arthur est le premier à parler. Il pense que Thomas est le meilleur candidat, et il le dit. Barbara, qui parle après Arthur, connaît maintenant l'opinion d'Arthur ; et si elle est, elle aussi, enthousiaste à propos de Thomas, il ne fait pas de doute qu'elle le soutiendra. Mais supposons que Barbara n'ait pas d'opinion arrêtée, et qu'elle ne soit pas certaine de savoir lequel des trois candidats est le meilleur. Si elle a confiance dans le jugement d'Arthur, elle sera tentée de suivre son avis : oui, Thomas est le meilleur. La confiance que lui inspire Arthur est suffisante pour qu'elle se rallie à son avis.

Après Arthur et Barbara, c'est maintenant au tour de Charles de s'exprimer. Son opinion n'est pas favorable à Thomas : selon lui, c'est Julie qui est la meilleure candidate. Pourtant, il est fort possible qu'il n'exprime pas cette opinion, et qu'il choisisse au contraire de suivre Arthur et Barbara en soutenant Thomas. Faut-il y voir de la lâcheté ? Pas nécessairement. D'abord, Charles sait que son opinion n'est fondée que sur des impressions et des informations imparfaites – nous sommes dans le domaine du jugement, pas dans celui des sciences exactes. Ensuite, Charles peut légitimement penser que l'enthousiasme d'Arthur et de Barbara repose sur

des éléments solides. En somme, Charles peut se dire que, s'il change d'avis, c'est parce qu'il sait écouter et tenir compte de l'opinion d'autrui.

La cascade informationnelle a un effet pervers qu'il faut souligner. Il est tout à fait possible que Charles ait recueilli des éléments d'information sur Thomas (ou sur les autres candidats) dont Arthur et Barbara n'ont pas connaissance. Ces informations, si elles avaient été mises sur la table, auraient peut-être pu faire changer Arthur ou Barbara d'avis. Ç'aurait été le cas si le hasard (ou le président de séance) avait désigné Charles pour parler en premier. Mais comme Charles a parlé après Arthur et Barbara, il a gardé ces informations pour lui.

On imagine aisément que cet effet va croissant avec le nombre de personnes qui expriment une opinion concordante. Plus celles qui ont annoncé leur préférence pour Thomas seront nombreuses, et plus on hésitera à s'opposer à elles sans de très solides raisons. En définitive, la prise de position initiale d'Arthur aura déclenché un processus de cascade, et le groupe choisira Thomas à l'unanimité – même si certains des participants n'avaient en réalité pas d'opinion ou préféraient un autre candidat.

Cet exemple est bien sûr très artificiel. Mais la mécanique des cascades informationnelles est présente dans tous les groupes. Elle rend possible (et même probable) que des groupes divergent de manière importante pour des raisons quasi aléatoires. Dans notre exemple, si Barbara, plutôt qu'Arthur, avait parlé la première et penché pour Sam plutôt que pour Thomas, ou si Arthur avait préféré Julie, on peut imaginer que le groupe aurait désigné Sam ou Julie, non pas parce qu'ils étaient meilleurs que Thomas, mais en raison de l'effet de cascade. C'est la même mécanique de cascades qui est à l'œuvre dans les téléchargements de chansons et dans les expériences similaires que nous avons mentionnées.

Il faut souligner qu'il n'est pas nécessairement irrationnel, du point de vue de chacun des acteurs, de participer à des cascades informationnelles.

Quand on n’est pas sûr soi-même de la personne à recruter, il n’est pas forcément idiot de suivre l’avis des autres. De manière générale, plus le nombre de personnes qui partagent la même opinion est élevé, et plus il est censé d’en tenir compte.

Ce raisonnement optimiste néglige toutefois deux problèmes. D’abord, la personne qui s’appuie sur l’opinion des autres n’a pas forcément conscience que ceux-ci sont eux-mêmes pris dans une cascade, et qu’ils n’expriment donc pas une opinion propre et indépendante. Dans notre exemple, Charles est impressionné par la convergence d’Arthur et de Barbara, mais ne songe pas que le premier a influencé la seconde. Plus généralement, quand nous observons trois, dix ou vingt personnes qui arrivent à la même conclusion, nous sous-estimons à quel point elles suivent toutes celles qui les ont précédées. Le consensus que nous interprétons comme le reflet de la sagesse collective peut n’être que le produit des opinions initiales de quelques personnes. Le second problème est empirique : les cascades informationnelles, on l’a vu, peuvent amener des groupes de grande taille à prendre de très mauvaises décisions. Après tout, il est possible qu’Arthur se soit complètement trompé sur Thomas.

Nous avons insisté ici sur les cascades informationnelles, mais elles ne sont bien sûr pas l’unique raison qui fait que les membres d’un groupe s’influencent mutuellement. La pression sociale joue également. Dans une entreprise ou dans une administration, il arrive que l’on choisisse de taire son désaccord, pour ne pas donner l’impression d’être négatif, borné, agressif ou antipathique. Chacun veut faire preuve d’esprit d’équipe, et pour cela, rien de tel que s’aligner sur les autres. On peut penser connaître la « bonne » décision, ou en tout cas savoir laquelle on préférerait, et choisir néanmoins de s’aligner sur le consensus apparent du groupe ou sur les premières opinions exprimées, simplement pour rester dans les bonnes grâces du groupe. On peut aussi choisir ses batailles : quand un consensus commence à se dégager, pourquoi se mettre en travers du chemin, alors que

la décision semble déjà pratiquement acquise ? Ainsi, dans notre parabole de la décision de recrutement, la pression sociale joue sans doute aussi un rôle : indépendamment des mérites supposés de Thomas, peut-être les participants ont-ils tout simplement choisi de « ne pas faire de vagues », afin d'éviter de paraître désagréables (ou idiots).

Le paradoxe de la cascade de pression sociale est identique à celui de la cascade informationnelle : on a tendance à surestimer la conviction des personnes qui se sont déjà exprimées, et à sous-estimer le rôle qu'a joué la pression sociale dans la formation de leur opinion. C'est ainsi qu'on ajoute sa voix au consensus, et qu'on accroît le niveau de pression sociale sur les participants suivants. Dans toutes les organisations, la conjonction des cascades informationnelles et des cascades de pression sociale peut produire une grande confiance, et un soutien unanime, sur un jugement totalement erroné.

La manière dont l'influence sociale amplifie des différences mineures et anecdotiques permet d'expliquer le bruit qu'on observe entre des groupes similaires. Quand un groupe apporte un soutien unanime à un projet, on oublie un peu vite que cet accord peut être le fruit non pas de sa conviction, mais des cascades informationnelles et de la pression sociale. Si quelqu'un d'autre avait parlé le premier pour exprimer une autre opinion, ou si le premier orateur avait décidé de garder le silence, les mêmes mécanismes auraient pu conduire le groupe dans une tout autre direction. C'est ainsi que des groupes très semblables et confrontés au même problème peuvent finir par prendre des décisions très différentes.

## *La polarisation de groupe*

Aux États-Unis et dans de nombreux autres pays, les affaires pénales (et certaines affaires civiles) sont généralement jugées par un jury. L'une des

raisons de recourir à des jurys populaires – ou, à tout le moins, l'une des raisons d'espérer que leurs décisions soient justes – est qu'on suppose que les jurys, grâce à leur diversité et à leurs délibérations, prennent collectivement des décisions plus sages que les individus qui les composent.

L'étude des jurys met pourtant en lumière un autre type d'influence sociale qui produit, lui aussi, du bruit : la *polarisation de groupe*. Le principe est simple : lorsque des gens discutent entre eux, ils finissent souvent par adopter une forme plus extrême de l'opinion qui était la leur au départ. Si, par exemple, un groupe de sept personnes doit décider d'ouvrir ou non un nouveau bureau à Paris, et si la plupart de ses participants ont plutôt tendance à penser que ce serait, tout bien considéré, une bonne idée, il est probable qu'après discussion, le groupe conclura qu'il s'agit d'une idée absolument brillante. La délibération a tendance à rendre les groupes plus unanimes et plus confiants dans leur jugement, tout en les amenant souvent à des positions plus extrêmes ou défendues avec plus d'enthousiasme.

Dans une série d'expériences, nous avons étudié les décisions prises par des jurys chargés d'indemniser des plaignants dans des affaires où des entreprises avaient commercialisé des produits défectueux. Chaque décision consistait donc à fixer le montant de dommages et intérêts approprié pour sanctionner l'entreprise coupable et en dissuader d'autres de l'imiter. Nous reviendrons en détail sur cette étude au chapitre 15, mais nous l'introduisons ici parce qu'elle permet d'illustrer la polarisation de groupe[7].

Pour commencer, nous avons présenté aux 899 participants de l'étude des descriptions simplifiées d'affaires réelles, et nous avons demandé à chacun, individuellement, de rendre un jugement sur chacune des affaires. Les participants devaient exprimer, sur une échelle de 1 à 7, le degré d'indignation que leur inspirait le comportement de l'entreprise fautive, ainsi que leur intention de la punir. Ils devaient également fixer, en dollars, le montant des dommages et intérêts (s'ils souhaitaient en accorder). Il nous

a ensuite suffi de traiter informatiquement ces réponses individuelles pour créer des millions de « jurys statistiques » : des groupes virtuels de six personnes choisies aléatoirement. Pour chacun de ces jurys virtuels, nous avons considéré la médiane des six jugements individuels comme le verdict du jury statistique.

Comme on pouvait s'y attendre, les jugements de tous ces jurys statistiques étaient remarquablement cohérents entre eux. Ce faible niveau de bruit entre jurys est un effet mécanique de l'agrégation statistique des jugements individuels au sein de chaque jury : le bruit présent dans des jugements individuels est fortement atténué quand on prend la médiane de six jugements indépendants.

Mais les jurys réels ne sont pas des jurys statistiques : ils se réunissent et délibèrent sur l'affaire. On peut raisonnablement se demander si un jury délibérant aura effectivement tendance à aboutir à un jugement proche de celui de son membre médian. Pour le savoir, nous avons fait suivre la première expérience d'une seconde, mobilisant cette fois plus de trois mille citoyens susceptibles d'être désignés comme jurés, que nous avons assemblés et fait délibérer dans plus de cinq cents jurys simulés[8].

Les résultats furent clairs et nets. Sur la même affaire, les jurys délibérants étaient bien plus sujets au bruit que les jurys statistiques, ce qui témoignait clairement du rôle de l'influence sociale dans la production de bruit. La délibération créait du bruit.

Un autre résultat retint notre attention. Dans un jury simulé de six personnes, quand le membre médian n'était que modérément indigné et prônait une sanction clémente, le verdict du jury, après délibération, était généralement encore plus clément. Quand, au contraire, le juré médian était absolument scandalisé et exprimait une ferme intention de punir l'entreprise fautive, le jury finissait généralement par être encore plus sévère que ce juré médian. La même tendance se retrouvait dans la fixation des dommages et intérêts : dès lors que le membre médian du jury proposait d'en accorder, le

jury avait systématiquement tendance à fixer un montant plus élevé que celui proposé par ce juré médian. 27 % des jurys fixaient même un montant d'indemnité supérieur ou égal à celui proposé par leur membre *le plus sévère*. En d'autres termes, non seulement les jurys délibérants étaient davantage sujets au bruit que les jurys statistiques, mais ils avaient tendance à accentuer, à rendre plus extrêmes les opinions des individus qui les composent.

Ces résultats illustrent parfaitement le principe de la polarisation de groupe : quand les gens se parlent, ils convergent en général sur une opinion plus radicale que leurs opinions initiales. La délibération conduit les jurys à une clémence plus grande quand le juré médian est clément, et à une sévérité plus grande quand le juré médian est sévère. De même, les jurys enclins à accorder des dommages et intérêts finissent par en fixer le montant à un niveau plus élevé que ne le proposent leurs membres médians.

Les raisons de la polarisation de groupe sont analogues à celles des effets de cascade. L'information y joue un rôle majeur. Dans un groupe dont la majorité des membres sont favorables à une peine sévère, on entendra davantage d'arguments en faveur de la sévérité que d'arguments contraires. Si, comme c'est probable, les membres du groupe s'écoutent mutuellement, leur opinion va évoluer dans le sens de la tendance dominante, renforçant ainsi la confiance et l'unité du groupe, et le rendant aussi plus radical. Et si les membres du groupe sont sensibles à la pression sociale, ceux qui sont, au départ, éloignés de la position médiane vont s'en rapprocher, ce qui produira là aussi de la polarisation.

La polarisation de groupe peut, bien sûr, produire des erreurs – et c'est souvent le cas. Mais ce qui nous intéresse d'abord ici, c'est la variabilité. Comme nous l'avons vu, l'agrégation de jugements multiples réduit le bruit, et ce d'autant plus que les jugements sont nombreux. C'est pourquoi il y a moins de bruit quand on compare les verdicts de jurys statistiques que quand on compare les jurés considérés individuellement. Pourtant, nous

avons vu aussi qu’il y a plus de bruit entre les jurys délibérants qu’entre les jurys statistiques. Quand on observe, comme c’est le cas ici, des différences importantes entre des groupes similaires qui traitent les mêmes problèmes, c’est souvent la polarisation de groupe qui en est responsable.

---

Partout où l’on décide en groupe, les effets de cascade et de polarisation peuvent conduire à de grandes disparités entre des groupes confrontés aux mêmes jugements. Les décisions finales peuvent dépendre fortement des jugements de quelques individus : ceux qui parlent en premier ou qui ont la plus grande influence. Comme les jugements de ces derniers sont eux-mêmes, nous l’avons vu, fortement sujets au bruit, les groupes délibérants, en dépit du cliché qui leur prête une sagesse supérieure, ont tendance à amplifier le bruit, et à y être plus sujets que les groupes « statistiques » définis par une simple moyenne des jugements individuels.

Pour toutes les organisations, où la majorité des décisions importantes sont prises à l’issue d’un processus délibératif, il est essentiel de prendre conscience de ce risque, afin de réduire le bruit, et non de l’amplifier. Les stratégies de réduction du bruit que nous allons proposer poursuivent notamment cet objectif.

## À propos des décisions de groupe

« Tout semble dépendre du succès initial. Nous avons donc intérêt à faire le maximum pour que notre nouveau produit fasse un malheur dès la première semaine. »

« Les idées en politique et en économie ressemblent beaucoup aux chansons ou aux stars de cinéma. Si chacun s'imagine que les autres apprécient une idée, alors elle peut aller très loin. »

« Nous avons décidé à l'unanimité ; mais sommes-nous sûrs que nous n'avons pas été victimes d'un effet de cascade ? »

« J'ai toujours été frappé de constater que, à l'issue de nos réunions d'équipe, nous sommes parfaitement confiants, unis et résolument engagés dans le sens de l'action que nous choisissons. Peut-être que nos délibérations ont tendance à nous polariser. Y a-t-il quelque chose à changer dans nos méthodes de discussion ? »

PARTIE III

# Le bruit dans les jugements prédictifs

Un grand nombre de jugements sont des prédictions. Et les prédictions, quand elles sont vérifiables, ont un grand avantage : on peut en évaluer l'exactitude. Leur étude a donc beaucoup à nous apprendre sur le biais et le bruit. Dans cette partie du livre, nous allons nous concentrer sur les jugements prédictifs.

Au chapitre 9, nous comparerons la justesse des prédictions faites par des professionnels, par des machines sophistiquées, et par des règles ou des formules simples. De manière générale, les professionnels arrivent bons derniers dans cette compétition. Au chapitre 10, nous nous pencherons sur les raisons de ce résultat et montrerons que le bruit est une très sérieuse limite à la qualité du jugement humain.

Pour pouvoir faire ces comparaisons, il faut évaluer la qualité des prédictions. Nous aurons besoin pour ce faire d'un moyen de mesurer leur justesse, c'est-à-dire de répondre à la question suivante : « Dans quelle mesure les prédictions *covarient*-elles avec les résultats ? » Si la direction des ressources humaines d'une entreprise, par exemple, fait une évaluation du potentiel des nouveaux embauchés, nous pouvons attendre quelques

années pour voir de quelle façon ils se sont comportés au travail et si l'évaluation de leur potentiel covarie avec celle de leurs résultats. Concrètement, les prédictions seront justes dans la mesure où les employés considérés comme « à fort potentiel » sont bien notés pour leur performance ; ou encore lorsque la liste des employés dont la performance laisse à désirer correspond peu ou prou à celle des nouveaux embauchés dont le potentiel avait été jugé limité.

Une mesure permet de quantifier cette correspondance de manière assez intuitive en répondant à une question plus précise. Nous l'appellerons le *pourcentage de concordance* (PC). Supposons que vous choisissiez deux employées au hasard : quelle est la probabilité que celle dont le potentiel a été le mieux noté au moment de l'embauche soit aussi celle qui a les meilleurs résultats dans son travail[1] ? Si la justesse des prévisions était parfaite, la réponse – le pourcentage de concordance – serait 100 % : le classement du potentiel serait une prévision parfaite du classement des performances. Si, à l'autre extrême, les prévisions de potentiel étaient parfaitement inutiles, la concordance serait purement aléatoire : l'employée qu'on avait jugée comme ayant le meilleur potentiel aurait une chance sur deux d'être en réalité la meilleure des deux. Le PC serait donc de 50 %.

Vous vous demandez peut-être quel PC on trouve en réalité dans cet exemple des décisions d'embauche. La réponse à cette question largement étudiée, que nous découvrirons au chapitre 9, vous surprendra sans doute. À ce stade, pour illustrer le principe du PC, un exemple plus simple fera l'affaire : le PC entre stature et longueur du pied chez l'homme adulte est de 71 %. Autrement dit, quand deux hommes rentrent dans un magasin de chaussures, il y a 71 % de chances que le plus grand des deux soit celui qui chausse la plus grande pointure.

Le PC a l'avantage de donner une mesure immédiate et intuitive du degré de covariation entre deux variables. Mais ce n'est pas l'indicateur habituellement utilisé en sciences sociales, où l'on préfère le *coefficient de*

*corrélation* (noté *r*). Le coefficient de corrélation varie entre 0 et 1 quand deux variables covarient dans le même sens (et entre 0 et –1 quand elles covarient en sens contraire). Dans l'exemple précédent, la corrélation entre stature et pointure est d'environ 0,60[2].

Il y a plusieurs façons de se représenter le coefficient de corrélation. L'une des plus intuitives est de voir la corrélation entre deux variables comme le pourcentage de leurs déterminants communs. Imaginons par exemple qu'un caractère soit déterminé de façon totalement génétique. Parmi les membres d'une même fratrie, qui ont en moyenne 50 % de gènes en commun, nous devrions trouver une corrélation de 0,50 sur ce caractère. Chez des cousins germains, qui ont 25 % de gènes en commun, la corrélation devrait être de 0,25. Inversement, quand nous observons une corrélation de 0,60 entre stature et pointure, nous pouvons en conclure que 60 % des facteurs de causalité qui déterminent la taille déterminent aussi la pointure.

Les deux mesures de covariation que nous venons de présenter sont directement liées. Le tableau 1 donne le PC pour différentes valeurs du coefficient de corrélation[3]. Dans la suite de ce livre, quand nous parlerons de la justesse d'un jugement prédictif, les deux mesures seront toujours présentées ensemble.

| Coefficient de corrélation (r) | Pourcentage de concordance (PC) |
|---|---|
| 0,0 | 50 % |
| 0,10 | 53 % |
| 0,20 | 56 % |
| 0,30 | 60 % |
| 0,40 | 63 % |
| 0,60 | 71 % |
| 0,80 | 79 % |
| 1,00 | 100 % |

**Tableau 1.** Coefficient de corrélation (r) et pourcentage de concordance (PC).

La justesse des jugements prédictifs souffre de notre sensibilité au bruit, mais aussi d'une autre limite, plus importante encore, que nous examinerons dans le chapitre 11. La plupart des jugements sont produits dans un état que nous appellerons l'*ignorance objective* : l'avenir dépend de beaucoup de choses que nous ne pouvons tout simplement pas connaître. Il est remarquable que nous réussissions la plupart du temps à oublier cette limite, et à formuler des prévisions avec confiance (et même avec une confiance excessive). Pour terminer, nous montrerons dans le chapitre 12 que l'ignorance objective n'affecte pas seulement notre capacité de *prévoir* les événements, mais aussi notre aptitude à les *comprendre*. Cette observation nous livrera l'une des clés d'une énigme : le fait que nous restions, la plupart du temps, sourds au bruit dans le jugement.

CHAPITRE 9

# Jugements et modèles

Nul ne s’en étonnera : toutes les organisations ont à cœur de prévoir quels individus seront performants dans leur travail. La prévision de la performance au travail a donc fait l’objet d’innombrables études, ce qui en fait un excellent exemple de jugement professionnel prédictif.

Pour l’illustrer, imaginons deux cadres dirigeantes d’une grande entreprise, Monica et Nathalie. Au moment de leur embauche, elles ont fait l’objet d’une évaluation approfondie par un cabinet spécialisé. Cette évaluation a été traduite en notes, sur une échelle de 1 à 10, sur différentes dimensions : le leadership, les capacités de communication, les compétences interpersonnelles, les compétences techniques liées au poste, et la motivation. Le tableau 2 présente les résultats. Le petit exercice auquel nous vous invitons consiste à prévoir, en utilisant également une échelle de 1 à 10, quelle sera l’évaluation des performances de ces deux cadres deux ans après leur embauche.

| | Leadership | Communication | Compétences interpersonnelles | Compétences techniques | Motivation | Votre prévision |
|---|---|---|---|---|---|---|
| **Monica** | 4 | 6 | 4 | 8 | 8 | |
| **Nathalie** | 8 | 10 | 6 | 7 | 6 | |

**Tableau 2.** Deux candidates pour un poste de direction.

Face à ce type de problème, la plupart des gens se contentent de jeter un œil à chacune des deux lignes et de porter un jugement rapide, en se livrant éventuellement à un rapide calcul mental de la note moyenne. Si vous avez procédé ainsi, vous avez

probablement conclu que Nathalie était la meilleure des deux candidates, et que la différence entre elle et Monica était de l'ordre de 1 à 2 points.

## *Jugement ou formule ?*

Il existe un terme technique pour désigner la démarche informelle que vous venez d'utiliser : le *jugement clinique*. Dans un jugement clinique, on considère les informations disponibles, parfois complétées par un calcul rapide, et on utilise son intuition pour produire un jugement. Le jugement « clinique » est, en fait, ce que nous appelons simplement « jugement » dans ce livre.

Supposons maintenant que vous vous soyez livré à cet exercice non seulement sur les profils de Monica et de Nathalie, mais que vous ayez pris part à une expérience plus vaste, où vous évaluez des centaines de profils comparables. Tous ces profils concernent des cadres qui ont été embauchés il y a quelques années, et qui ont été, à l'époque, évalués sur les cinq mêmes dimensions. Sur la base de ces évaluations, vous avez utilisé votre jugement « clinique » pour formuler une prévision quant à leur réussite professionnelle future, comme vous venez de le faire en produisant une évaluation globale des chances de Nathalie et Monica. Les évaluations des performances de tous ces cadres dans leurs nouveaux rôles sont désormais disponibles. Dans quelle mesure ces évaluations de performance correspondent-elles à vos jugements cliniques sur leur potentiel ?

Cette question s'inspire (en la simplifiant) d'une étude réelle sur la prévision de la performance[1]. Il est fort probable que, si vous aviez fait partie des sujets de cette étude, vous ne seriez pas fier du résultat. En l'occurrence, les participants étaient des professionnels des ressources humaines, tous titulaires de doctorats en psychologie, employés par un cabinet international de consultants spécialisé dans la sélection et l'évaluation. La corrélation entre les prédictions de potentiel réalisées par ces experts et la performance réellement mesurée était de 0,15 (PC = 55 %). Autrement dit, quand ils donnaient une meilleure note à un candidat potentiel qu'à un autre (comme vous l'avez fait en préférant, par exemple, Nathalie à Monica), la probabilité que leur candidat favori obtienne par la suite une meilleure évaluation de performance était de 55 %. À peine mieux que s'ils avaient tiré à pile ou face ! Le moins qu'on puisse dire est que ces experts n'impressionnent pas par leur clairvoyance.

Mais peut-être pensez-vous que la limite de ces jugements cliniques tient aux notes qui leur servent de base. Si ces notes n'étaient d'aucune valeur prédictive, il ne serait pas étonnant qu'il en aille de même des appréciations globales des psychologues. Nous devons donc nous poser ces questions : combien ces notes contiennent-elles d'informations prédictives valides ? Comment peut-on les combiner pour obtenir la meilleure prédiction globale possible, c'est-à-dire celle qui aura la corrélation la plus élevée possible avec la performance ?

Il existe une méthode statistique classique pour répondre à ce genre de question. Dans l'étude présente, elle fournit une corrélation optimale de 0,32 (PC = 60 %). C'est certes toujours loin d'être impressionnant ; mais c'est quand même beaucoup mieux que les résultats des prévisions cliniques.

Dans cette technique, appelée *régression linéaire multiple* ou simplement *régression multiple*, la prédiction est une moyenne pondérée des données disponibles, appelées *prédicteurs* : chaque prédicteur est affecté d'un coefficient qui lui donne un poids plus ou moins grand dans la prédiction finale (un peu comme la moyenne générale à un examen est la moyenne des notes dans les différentes matières, pondérées par des coefficients différents[2]). Ce que la régression multiple permet, c'est de trouver les coefficients optimaux pour cette formule, c'est-à-dire les coefficients qui donnent la meilleure prévision possible. Par « meilleure prévision possible », on entend celle pour laquelle la corrélation entre la prévision et la variable cible est la plus élevée possible. (C'est aussi celle qui minimise l'erreur quadratique moyenne des prédictions, ce qui illustre bien le rôle dominant du principe des moindres carrés en statistique.)

Comme on peut s'y attendre, dans la formule de prévision optimale, le prédicteur qui est le plus étroitement corrélé avec la variable cible sera affecté du coefficient le plus important, et s'il existe des « prédicteurs » totalement inutiles (c'est-à-dire qu'ils ne prédisent en réalité rien du tout), ceux-ci seront multipliés par un coefficient zéro[3]. Les coefficients peuvent aussi être négatifs : dans notre exemple de prédiction de la performance managériale, si l'on connaissait le nombre d'amendes impayées ou de redressements fiscaux accumulés par les candidats, ces prédicteurs seraient peut-être affectés d'un coefficient négatif.

L'utilisation de la régression multiple est un exemple de *prédiction statistique*, également appelée *prédiction actuarielle* ou simplement *prédiction mécanique*. Il existe de nombreux types de prédiction mécanique, qui vont de règles simples (« recruter toute personne ayant le bac ») à des modèles d'intelligence artificielle sophistiqués. Mais les modèles de régression multiple sont considérés, pour citer une étude de référence,

comme « l'outil dominant dans la recherche sur le jugement et la prise de décision[4] »). Pour minimiser le jargon, nous appellerons ces modèles des *modèles simples*.

L'étude qui met en scène « Monica et Nathalie » est représentative d'un grand nombre d'études comparant la qualité des prédictions cliniques et des prédictions mécaniques. Toutes ces études ont en commun une structure simple[5] :

- Elles utilisent un ensemble de variables indépendantes comme *prédicteurs* (ici, les notes des candidats) pour prévoir une variable dépendante, que nous appellerons plus simplement le *résultat cible* (ici, les évaluations de la performance des mêmes individus).
- Des juges humains font des *prévisions cliniques*.
- Une règle (comme la régression multiple) utilise les mêmes prédicteurs pour produire des *prévisions mécaniques* des mêmes résultats cibles.
- On compare la justesse globale des prédictions cliniques et celle des prédictions mécaniques.

## *Meehl : le modèle optimal fait mieux que vous*

Lorsqu'on présente à des profanes cette structure d'étude, la question qui leur brûle les lèvres est bien sûr toujours la même : de la prédiction clinique ou de la prédiction mécanique, laquelle marche le mieux ? Que vaut le jugement humain à côté d'une formule ?

La question est sans doute ancienne, mais elle n'avait pas suscité beaucoup d'intérêt avant 1954, date à laquelle Paul Meehl, professeur de psychologie à l'Université du Minnesota, publia un livre, *Prédiction clinique et prédiction statistique : analyse théorique et examen des preuves*[6]. Meehl y faisait la synthèse d'une vingtaine d'études dans lesquelles un jugement clinique était comparé à une prédiction mécanique, sur des problèmes allant de la réussite scolaire au pronostic psychiatrique. Meehl y concluait sans ambiguïté que les règles automatiques étaient généralement supérieures au jugement humain. Ce faisant, il suggérait aussi que les cliniciens et autres professionnels

du jugement faisaient montre d'une faiblesse inquiétante là précisément où ils croyaient que résidait leur principale force : la capacité à synthétiser des informations multiples.

Pour mesurer à quel point cette découverte est surprenante, et quel est son lien avec le bruit, il faut bien comprendre comment fonctionne un modèle simple de prédiction mécanique. La caractéristique principale d'une prédiction mécanique est qu'elle applique la même règle, de manière uniforme et aveugle, à tous les cas. Chaque prédicteur est affecté d'un coefficient, et ce coefficient ne varie pas d'un individu à l'autre.

On pourrait penser que cette rigidité désavantage sévèrement le modèle par rapport au juge humain. Dans le cas de Monica, par exemple, peut-être avez-vous pensé, en examinant son profil, que la conjonction de sa forte motivation et de ses excellentes compétences techniques lui permettrait de compenser ses limites dans d'autres domaines. Inversement, peut-être vous êtes-vous dit que les faiblesses de Nathalie dans ces deux domaines n'étaient pas très graves au vu de ses autres points forts. Implicitement, vous avez imaginé différentes trajectoires de réussite pour les deux femmes. De fait, ces spéculations cliniques, qui semblent entièrement plausibles, reviennent à attribuer des poids différents aux mêmes prédicteurs dans le cas de Monica et dans celui de Nathalie. Une telle subtilité est hors de portée d'un modèle simple.

Une autre restriction de tout modèle simple est qu'une augmentation d'une unité pour un prédicteur donné produit toujours le même effet (et produit précisément la moitié de l'effet que produirait une augmentation de 2 unités). L'intuition clinique s'affranchit souvent de cette règle. Si, par exemple, la note maximale de 10 obtenue par Nathalie en communication vous a fait forte impression, et si cela lui a valu un « bonus » dans votre prévision, vous avez fait quelque chose qu'un modèle simple ne fera jamais. Aux « yeux » d'une formule qui calcule une moyenne pondérée, la différence entre un 10 et un 9 est exactement la même qu'entre un 7 et un 6. Le jugement clinique n'est pas soumis à cette contrainte : l'intuition suggère qu'un même écart peut être critique dans un certain contexte et sans conséquence dans un autre. Si vous avez intuitivement attribué une note globale à Nathalie et à Monica, vous pouvez vérifier s'il existe un modèle simple qui répliquerait exactement vos jugements : la réponse est probablement négative, parce que votre jugement clinique ne se soumet pas au carcan qu'impose une formule unique.

L'étude que nous avons décrite ici fournit un exemple clair du schéma identifié par Meehl. Comme nous l'avons noté, la corrélation avec la performance qu'obtenaient les prédictions cliniques était de 0,15 (PC = 55 %), tandis que la prédiction mécanique

atteignait 0,32 (PC = 60 %). Songez à la confiance que vous avez ressentie en évaluant les mérites relatifs de Monica et de Nathalie. Ces chiffres, et plus généralement les résultats des travaux de Meehl, suggèrent fortement que la confiance que vous avez éprouvée dans la qualité de votre jugement était une illusion : l'*illusion de validité*.

L'illusion de validité est présente partout où l'on produit des jugements prédictifs. Ce qu'elle reflète, c'est une confusion systématique entre deux étapes de la tâche de prédiction : d'abord, l'*évaluation* des cas en fonction des éléments disponibles ; ensuite, la *prédiction* des résultats réels. Il est bien sûr possible d'évaluer avec confiance laquelle des deux candidates *paraît* la meilleure au vu des éléments dont vous disposez. Mais laquelle *sera* réellement meilleure lorsque ses performances seront mesurées ? C'est une tout autre histoire ! Il est assez clair, par exemple, que Nathalie semble être une meilleure candidate que Monica, mais il n'en demeure pas moins très risqué d'affirmer qu'elle réussira mieux que Monica comme cadre dirigeante. La raison en est simple : même si les informations dont vous disposez sont très pertinentes pour évaluer les deux profils aujourd'hui, l'avenir est nécessairement incertain.

Si cette distinction entre les cas et les prédictions vous laisse perplexe[7], vous n'êtes pas seul. La distinction est difficile pour tout le monde. Mais si vous avez la même confiance dans vos prévisions que dans votre évaluation des cas, vous êtes victime de l'illusion de validité.

Les cliniciens ne sont pas à l'abri de cette illusion. Vous pouvez certainement imaginer la réaction des psychologues cliniciens aux travaux de Meehl. L'idée que des formules simples et appliquées de façon systématique puissent surpasser le jugement clinique suscita non seulement de la surprise, mais aussi beaucoup d'incrédulité, voire de mépris. Comment des études superficielles, utilisant des formules simplistes, osaient-elles se prononcer sur les merveilles de l'intuition clinique ? Le schéma de Meehl contredit l'expérience subjective du jugement, et la plupart d'entre nous avons tendance à faire davantage confiance à notre expérience qu'à la thèse d'un universitaire.

Signe de cette tension, le regard de Meehl sur ses propres travaux n'était pas dénué d'ambivalence. Comme le nom de Meehl est associé à la supériorité des statistiques sur le jugement clinique, il est tentant de l'imaginer en pourfendeur impitoyable de l'intuition humaine ou en saint patron des modélisateurs. Mais la caricature raterait son sujet. En plus d'être professeur d'université, Meehl était psychanalyste. Une photographie de Freud trônait dans son bureau[8]. C'était un touche-à-tout, qui enseignait la psychologie, mais aussi le droit et la philosophie[9]. Il est l'auteur de textes sur la métaphysique, la science politique, la religion et même la parapsychologie (il était

convaincu qu'il y avait « quelque chose dans la télépathie »). Rien de tout cela ne correspond au stéréotype du matheux qui ne jure que par les chiffres. Meehl n'avait aucun *a priori* contre les cliniciens, bien au contraire. Mais les preuves de la supériorité de l'approche mécanique étaient, écrivit-il, « massives et concordantes [10] ».

« Massives et concordantes » est un bon résumé. En 2000, une synthèse des résultats de cent trente-six expériences différentes confirma sans ambiguïté que l'agrégation mécanique surpasse le jugement clinique [11]. Les travaux analysés couvraient une grande diversité de sujets, parmi lesquels le diagnostic de la jaunisse, l'aptitude physique au service militaire et la satisfaction conjugale. La prédiction mécanique était plus exacte dans soixante-trois études, tandis que la prédiction clinique ne l'emportait que dans huit cas (un « match nul » statistique étant déclaré dans les soixante-cinq études restantes). Encore ces résultats sous-estiment-ils les avantages de la prédiction mécanique, puisque celle-ci est aussi plus rapide et moins coûteuse que le jugement clinique. En outre, les juges humains jouissaient en réalité d'un avantage indu dans beaucoup de ces études : ayant collecté eux-mêmes une partie des données, ils avaient accès à des informations « privées » qui n'étaient pas codées pour être fournies au modèle statistique [12]. La somme de ces études suggère une conclusion sans appel : *les modèles simples font mieux que les humains*.

## *Goldberg : un modèle de vous fait mieux que vous*

La découverte de Meehl soulève des questions importantes. Pourquoi, exactement, la formule est-elle supérieure à l'homme ? Qu'est-ce qu'elle fait mieux ?

Comme nous allons le voir, il vaudrait mieux demander *ce que les humains font moins bien*. De fait, ils sont inférieurs aux modèles statistiques de plusieurs manières, et l'une des principales est le bruit dans leurs jugements.

Pour le comprendre, nous devons nous intéresser à un autre courant de recherche sur les modèles simples, né dans la petite ville d'Eugene, dans l'Oregon. C'est là que Paul Hoffman, un psychologue visionnaire et fortuné, las des lenteurs et des rigidités du milieu universitaire, créa son propre institut de recherche. Il y réunit quelques chercheurs exceptionnels, qui firent d'Eugene un des plus grands centres mondiaux de l'étude du jugement humain.

Parmi ces chercheurs se trouvait Lewis Goldberg, qui est notamment passé à la postérité pour son rôle déterminant dans le développement du modèle de personnalité en cinq facteurs (les Big Five). À la fin des années 1960, en s'appuyant sur de premiers travaux d'Hoffman, Goldberg étudia les modèles statistiques qui permettent de décrire les jugements d'un individu [13].

Il est tout aussi facile de modéliser un individu (un « juge », dans notre terminologie) que de modéliser la réalité. On utilise les mêmes prédicteurs – dans notre exemple initial, ce seraient les cinq notes données aux candidates. On utilise aussi le même outil : la régression multiple. La seule différence est le résultat cible. Au lieu de prédire les performances *réelles* des candidates, la formule qu'on calcule doit prédire au mieux l'ensemble des *jugements* d'un juge donné – par exemple *vos* jugements sur Nathalie, Monica et d'autres cadres.

L'idée de modéliser vos jugements sous forme de moyenne pondérée peut sembler étrange. À l'évidence, ce n'est pas ainsi que vous vous formez vos opinions. Lorsque vous réfléchissez « cliniquement » aux cas de Nathalie et de Monica, vous ne leur appliquez pas la même formule. D'ailleurs, vous ne leur appliquez pas de formule du tout. Le modèle du juge n'est pas une description réaliste de la manière dont un juge produit un jugement.

Cependant, même si vous ne calculez pas vraiment une moyenne pondérée, rien n'interdit de supposer que *tout se passe comme si* vous la calculiez. Un excellent joueur de billard, par exemple, agit sensiblement *comme si* il avait résolu les équations complexes qui décrivent la mécanique de tel ou tel coup, bien qu'il ne fasse en réalité rien de tel [14]. De la même façon, vos prédictions peuvent ressembler à ce que vous obtiendriez en utilisant une formule simple – même si votre réflexion est beaucoup plus riche. Dans la mesure où un modèle « comme si » prédit avec une justesse raisonnable ce que va faire un juge, ce modèle est utile à notre compréhension, même s'il ne décrit évidemment pas la manière dont le juge raisonne.

Il s'avère que les modèles simples du jugement satisfont à ce critère. Une analyse de deux cent trente-sept études sur le jugement a montré que la corrélation moyenne entre un « modèle du juge » et les jugements cliniques produits par le juge en question était de 0,80 (PC = 79 %). Bien qu'imparfaite, cette corrélation est suffisamment élevée pour étayer une théorie du « tout se passe comme si [15] ».

Une fois admis le principe de modéliser le juge, la question centrale de Goldberg était de savoir comment un simple modèle du juge prédirait non seulement le juge, mais surtout les résultats *réels*. Comme le modèle est une grossière approximation du juge, on

s’attendrait naturellement à ce que ses résultats ne soient pas aussi bons. Mais de combien ? Que perd-on en exactitude quand le modèle du juge remplace le juge ?

La réponse vous surprendra sans doute. Goldberg découvrit que, quand on remplace le juge par un modèle du juge, les prédictions ne perdent pas en justesse, mais, au contraire, s’améliorent. Dans la plupart des cas, le modèle fait de meilleures prédictions que le professionnel sur lequel il était fondé. L’ersatz est meilleur que l’original.

Cette conclusion a été confirmée par des études dans un grand nombre de disciplines. Une des premières réplications du travail de Goldberg portait sur la réussite d’étudiants à l’université[16]. Les chercheurs demandèrent à 98 participants de prévoir, en utilisant dix prédicteurs, la note moyenne générale (*grade point average* ou GPA) qu’obtiendraient 90 étudiants. Sur la base de ces prédictions, les chercheurs développèrent ensuite des modèles linéaires du jugement de chacun des participants, créant ainsi 90 « modèles de juges ». Enfin, ils comparèrent la justesse des prédictions de chaque participant à celle des prédictions de « son » modèle. Pour chacun des 90 participants, le modèle faisait mieux ! Bien des années plus tard, une synthèse portant sur cinquante ans de recherche confirma que les « modèles de juges » font systématiquement de meilleures prévisions que les juges qu’ils modélisent[17].

Nous ne savons pas si les participants à ces études ont reçu un feed-back personnel sur leur performance. Mais vous pouvez certainement imaginer votre propre désarroi si l’on vous expliquait qu’une formule qui n’est dérivée d’aucune autre information que vos jugements, et qui en propose un modèle grossier, voire caricatural, parvient en réalité à voir plus juste que vous. Pour la plupart d’entre nous, si le jugement est une activité complexe, intéressante et riche, c’est précisément parce qu’elle n’obéit *pas* à des règles simples. Nous sommes particulièrement contents de nous et de notre aptitude à produire des jugements quand nous imaginons des raisonnements complexes, ou quand nous avons une intuition qui différencie un cas individuel des autres. Si le jugement humain – et le nôtre en particulier – nous semble précieux, c’est justement quand les jugements sont subtils, donc irréductibles à une opération aussi élémentaire qu’une moyenne pondérée. Les études sur le modèle du juge mettent à mal cette intuition, et confortent la conclusion de Meehl : toute cette subtilité est vaine. La complexité et la richesse du raisonnement ne produisent généralement pas de meilleures prédictions.

Quelle en est la raison ? Pour comprendre les résultats de Goldberg, il nous faut analyser l’écart entre vous et le « modèle de vous ». D’où viennent les disparités entre vos jugements réels et ceux produits par un modèle simple qui tente de les prédire ?

Répétons-le : un modèle statistique de vos jugements n'ajoute rien aux informations que ces jugements contiennent. Il ne sait rien que vous ne sachiez déjà. Tout ce que le modèle peut faire, c'est soustraire et simplifier. En particulier, un modèle simple de vos jugements ne pourra saisir aucune règle complexe, même si c'est une règle que vous suivez systématiquement. Si, par exemple, vous pensez que la différence entre un 10 et un 9 pour la note en matière de communication est plus importante que la différence entre un 7 et un 6, vous appliquez une règle complexe, qu'une formule linéaire ne saurait refléter. Il en va de même si vous pensez qu'un profil « harmonieux », qui obtient la même note de 7 sur chacune des dimensions considérées, est supérieur (ou inférieur) à un candidat « contrasté » qui obtient la même moyenne avec des points forts et des points faibles marqués. Le « modèle de vous » ne comprendra pas et ne reproduira pas ces règles complexes – même si vous les appliquez avec une cohérence sans faille.

Le fait de ne pas reproduire ces règles complexes et subtiles se traduira nécessairement par une perte d'exactitude – du moins, lorsque votre subtilité est réelle, et votre règle pertinente. Supposons, par exemple, que vous deviez prédire la réussite d'un individu dans une tâche difficile à partir de deux facteurs : la compétence et la motivation. Aucune moyenne pondérée ne donnera sans doute une très bonne formule, car même une excellente motivation ne suffit pas à compenser un grave déficit de compétence, et *vice versa*. Si vous vous servez intuitivement d'une combinaison plus complexe des deux facteurs, votre justesse prédictive en sera accrue. Elle sera aussi plus élevée que celle obtenue par un modèle linéaire incapable d'appréhender cette subtilité. Mais bien sûr, dans cet exemple, la règle complexe semble pertinente. En réalité, vous pourriez aussi imaginer toutes sortes de règles de jugement complexes qui ne font que nourrir l'illusion de validité et nuisent à la qualité des jugements. Certaines subtilités sont pertinentes, mais beaucoup ne le sont pas.

Outre la présence de règles complexes, il existe une autre source d'écart entre vous et le « modèle de vous » : le bruit. Le modèle ne reproduira pas ce que nous avons appelé le bruit de pattern de vos jugements. Nous avons vu que le bruit de pattern résulte pour partie des erreurs positives et négatives dues à vos réactions arbitraires à des cas particuliers. Par construction, le modèle de vous suppose que vous êtes cohérent d'un cas à l'autre : il ne peut pas répliquer ces jugements au cas par cas. Le modèle n'appréhendera pas non plus l'influence du contexte dans lequel vous produisez un jugement, ni celle de votre état d'esprit à un instant donné. Ces sources de bruit occasionnel ne sont pas corrélées avec quoi que ce soit d'autre, ce qui signifie qu'elles peuvent être considérées, en pratique, comme aléatoires.

La suppression du bruit qui affecte vos jugements aura toujours pour effet d'améliorer la justesse des prédictions[18]. Il s'agit là d'une simple loi statistique. Pour l'illustrer, supposons par exemple que la corrélation entre vos prévisions et le résultat cible soit de 0,50 (PC = 67 %), mais que la variance de vos jugements se compose pour moitié de bruit. Si vos jugements pouvaient être exempts de tout bruit – comme le serait un modèle de vous – leur corrélation avec le même résultat bondirait à 0,71 (PC = 75 %). Réduire le bruit accroît automatiquement la validité du jugement prédictif.

En somme, en vous remplaçant par un « modèle de vous », on élimine deux choses : d'une part, votre subtilité ; d'autre part, votre bruit de pattern (stable ou occasionnel). Or, comme nous l'avons vu, l'observation empirique nous enseigne que le modèle du juge est plus valide que le juge sur lequel le modèle se fonde. Le rapprochement de ces deux observations suggère une conclusion capitale : les gains qu'apportent des règles subtiles dans le jugement humain – si ces règles et ces gains existent – ne sont généralement pas suffisants pour compenser les effets préjudiciables du bruit. Comparé à une formule qui vous réduit à une caricature de votre façon de penser, vous avez certainement l'impression d'être plus subtil, plus perspicace et plus nuancé. Mais en réalité, vous produisez surtout beaucoup plus de bruit.

Pourquoi des règles complexes nuisent-elles à la justesse de nos prédictions, alors que nous sommes convaincus qu'elles reposent sur des intuitions pertinentes ? D'abord, cette conviction peut être erronée : toutes les règles complexes qu'inventent les juges face à un problème donné ne sont pas forcément justes. Mais ce n'est pas tout. Même quand une règle complexe est valide en principe, elle s'applique inévitablement dans des conditions qui ne sont que rarement observées.

Supposons par exemple que vous ayez tiré de votre expérience et de votre intuition cette leçon : les candidats qui font preuve d'une exceptionnelle originalité font toujours d'excellentes recrues, même quand ils sont médiocrement notés dans d'autres domaines. Le problème est que les candidats exceptionnellement originaux sont par définition exceptionnellement rares. Comme, de surcroît, votre évaluation de l'originalité n'est probablement pas très fiable, beaucoup des notes élevées sur ce critère peuvent être données par accident (et des candidats réellement originaux passer inaperçus). Quant aux évaluations de performance à la lumière desquelles vous avez conclu que les individus « originaux » deviennent un jour des « superstars », elles sont également imparfaites. Ces erreurs de mesure, tant sur l'évaluation de l'originalité que sur celle de la performance, atténuent inévitablement la justesse des prédictions, surtout lorsqu'il

s’agit d’événements rares. Les avantages de la véritable subtilité sont rapidement noyés dans l’erreur de mesure.

Une étude de Martin Yu et Nathan Kuncel propose une version plus radicale encore de la démonstration de Goldberg. Cette étude, dont s’inspire l’exemple de Monica et Nathalie, utilisait les données d’un cabinet de consultants international, dont les experts avaient évalué 847 candidats à des fonctions de cadre dirigeant, dans trois échantillons séparés[19]. Les notes reçues par les candidats sur sept dimensions distinctes avaient servi de base aux experts pour attribuer, en utilisant leur jugement clinique, une note globale à chacun, avec une qualité prédictive loin d’être impressionnante.

Yu et Kuncel décidèrent d’aller au-delà de la comparaison entre les juges et les « modèles du juge », en ayant recours à des modèles linéaires *aléatoires*. Pour ce faire, les chercheurs générèrent dix mille modèles différents, chacun affectant aux sept prédicteurs un ensemble de coefficients déterminé par le hasard. Ils testèrent ensuite les dix mille formules ainsi générées pour savoir si certains de ces modèles simples parviendraient à prédire convenablement la performance réelle des candidats[20].

Leur résultat le plus étonnant était que pratiquement *n’importe quel* modèle simple, pourvu qu’il soit appliqué à tous les cas avec systématisme, fait mieux que les juges humains utilisant les mêmes informations. Dans un des trois échantillons, 77 % des dix mille modèles linéaires aléatoirement pondérés faisaient mieux que les experts humains ; dans les deux autres, c’était 100 %. Ou, pour le dire simplement : il s’avéra presque impossible, dans cette étude, de produire un modèle simple qui fasse *moins bien* que les experts.

La leçon de cette étude est plus frappante encore que celle que nous avons tirée du travail de Goldberg sur le modèle du juge. Elle constitue, il est vrai, un exemple extrême : dans ce cas particulier, les juges humains avaient de très mauvais résultats en termes absolus, ce qui aide à comprendre pourquoi même des modèles linéaires générés au hasard peuvent faire mieux. Il serait bien sûr erroné de conclure que n’importe quel modèle, en toutes circonstances, fait mieux que n’importe quel humain. Il n’empêche : l’adhésion automatique à une règle simple, même quand cette règle est loin d’être optimale, peut grandement améliorer le jugement dans un problème difficile. L’expression qu’emploient Yu et Kuncel est « constance irréfléchie ». Le gain qu’apporte la constance, même irréfléchie, démontre en creux l’effet négatif massif qu’a l’inconstance – le bruit – sur la validité des prédictions cliniques.

Cet exposé rapide a montré l'effet préjudiciable du bruit sur la qualité du jugement clinique. Dans les jugements prédictifs, les experts humains sont facilement surpassés par des formules simples, qu'il s'agisse de modèles de la réalité, de « modèles du juge », ou même de modèles générés de façon aléatoire. Cette découverte plaide en faveur de l'utilisation de méthodes exemptes de bruit : les règles et les algorithmes. C'est le sujet du prochain chapitre.

### À propos des jugements et des modèles

« Plus de soixante ans après la publication du livre de Paul Meehl, l'idée que la prédiction mécanique est supérieure au jugement humain étonne encore. »
« Les gens croient appréhender la complexité et ajouter de la subtilité quand ils produisent des jugements. Mais la complexité et la subtilité améliorent rarement la justesse des modèles simples. »
« Il y a tellement de bruit dans le jugement qu'un modèle du juge, simplement parce qu'il est exempt de bruit, fait des prédictions plus justes que le juge dont il s'inspire. »
« Même un modèle aléatoire peut faire mieux qu'un juge humain : la constance irréfléchie fait souvent mieux que la réflexion inconstante. »

CHAPITRE 10

# Le silence des formules

Nous nous sommes désormais habitués à voir l'intelligence artificielle (IA) réaliser de nombreuses tâches considérées jusque-là comme l'apanage des humains. Les algorithmes de *machine learning* (« apprentissage automatique ») peuvent reconnaître des visages, traduire des langues, interpréter des radiographies. Ils peuvent résoudre des problèmes computationnels, par exemple pour donner des indications d'itinéraire à des milliers de conducteurs en même temps, avec une précision et une rapidité stupéfiantes. Et ils excellent aussi dans des tâches de prédiction complexes : prévoir les décisions de la Cour suprême des États-Unis, identifier les prévenus ayant la plus forte probabilité de manquer à leur obligation de comparaître, ou encore déterminer quels signalements aux services de protection de l'enfance nécessitent la visite urgente d'un travailleur social.

Si ce sont aujourd'hui les applications qui nous viennent à l'esprit quand nous entendons le mot *algorithme*, le terme est plus ancien et a une signification plus large. Un dictionnaire les définit comme des « processus ou ensembles de règles devant être suivis pour réaliser des calculs ou d'autres opérations de résolution de problème, en particulier par un ordinateur ». En ce sens, les modèles simples et les autres formes de

prédiction mécanique que nous avons décrits dans le précédent chapitre sont aussi des algorithmes.

En réalité, de nombreuses méthodes de prédiction mécanique, allant de règles ridiculement simples à des algorithmes d'une sophistication extrême, peuvent surpasser le jugement humain. Et l'une des principales raisons de cette supériorité est que toute méthode mécanique est, par construction, exempte de bruit.

Pour examiner ces différentes méthodes et comprendre l'intérêt de chacune d'entre elles, nous commencerons notre exploration par les modèles de régression linéaire, c'est-à-dire les « modèles simples » du chapitre 9. À partir de là, nous prendrons deux directions opposées : d'une part pour rechercher la simplicité extrême, avec des modèles dits « frugaux », d'autre part pour ajouter plus de raffinement et de complexité, avec les outils modernes de l'intelligence artificielle (figure 11).

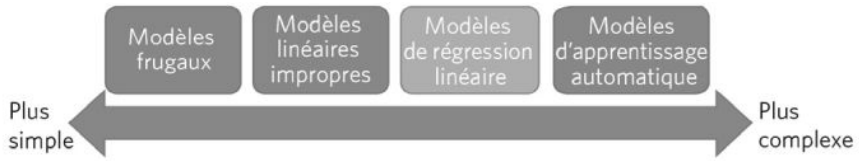



**Figure 11.** Quatre types de règles et d'algorithmes.

## *Vers plus de simplicité : la beauté dans la robustesse*

Robyn Dawes faisait partie des stars de la recherche réunies dans les années 1960 et 1970 à Eugene, dans l'Oregon, pour étudier le jugement. En 1974, Dawes fit une découverte capitale en matière de simplification des tâches prédictives. Son idée était originale, voire hérétique : au lieu

d'utiliser la régression multiple pour déterminer précisément le coefficient idéal dont chaque prédicteur devait être affecté, il proposait de donner à chaque prédicteur un poids égal.

Ce que Dawes découvrit, et qui ne manqua pas d'étonner ses contemporains, c'est qu'un tel *modèle équipondéré* (donnant à chaque prédicteur le même poids) est à peu près aussi performant qu'un modèle de régression dont les coefficients sont optimisés par la méthode classique, et très supérieur au jugement clinique[1]. Dawes donna à ce type de modèle, et plus généralement aux modèles linéaires qui n'utilisent pas les coefficients optimaux, le nom de *modèles linéaires impropres*.

Même les partisans des modèles impropres reconnaissent que cette idée est choquante et « contraire à l'intuition statistique[2] ». Dawes et son assistant Bernard Corrigan eurent d'ailleurs toutes les peines du monde à publier leur article initial sur le sujet : les éditeurs de revues scientifiques ne les croyaient tout simplement pas. Songez au cas de Nathalie et Monica évoqué dans le précédent chapitre : il vous semble probablement évident que certains prédicteurs sont plus pertinents que d'autres. S'agissant de cadres dirigeants, par exemple, la plupart des gens accordent aux qualités de leadership un poids plus grand qu'aux compétences techniques. Comment la simple moyenne non pondérée de ces notes pourrait-elle prédire la performance d'un individu avec plus d'exactitude qu'une moyenne soigneusement pondérée, ou que le jugement d'un expert ?

Aujourd'hui, bien des années après la découverte de Robyn Dawes, le phénomène statistique qui laissait ses contemporains perplexes ne surprend plus. Il a en effet une explication purement technique. Nous la résumons ici pour les lecteurs intéressés : ceux que les mystères des statistiques ne fascinent pas n'hésiteront pas à sauter à la section suivante.

Comme nous l'avons vu au chapitre précédent, la régression multiple calcule les coefficients « optimaux », c'est-à-dire ceux qui minimisent l'erreur quadratique. Mais la régression multiple minimise l'erreur *dans*

*l'échantillon de départ* qui sert de base à son calcul. Elle traite en quelque sorte cet échantillon comme s'il représentait la totalité du monde, et en tire toutes les leçons possibles. Le danger est donc qu'elle en tire, en fait, *trop* de leçons, en s'ajustant à tous les cas extrêmes que le hasard a placés dans l'échantillon de données sur lequel elle est calculée.

Supposons par exemple que l'échantillon comprenne quelques cadres dotés de compétences techniques élevées, et qui – sans que cela ait de rapport direct avec les compétences en question – ont obtenu des résultats exceptionnels. Le modèle en conclura que les compétences techniques élevées sont associées à une performance exceptionnelle. Il affectera donc le prédicteur « compétences techniques » d'un coefficient élevé.

Il est facile de comprendre comment une telle démarche conduit à des erreurs. Quand le modèle optimisé pour un échantillon de départ sera appliqué *hors échantillon*, c'est-à-dire pour prédire la performance future d'un *autre* ensemble de cadres, que se passera-t-il ? Les cas extrêmes qui étaient présents par hasard dans l'échantillon original ne le seront plus, précisément parce qu'ils n'étaient que des hasards : dans le nouvel échantillon, les cadres dotés de compétences techniques élevées ne seront pas tous des superstars. Inversement, le hasard placera dans le nouvel échantillon des cas extrêmes différents, que le modèle ne saura pas prédire.

Pour cette raison, la mesure correcte de la justesse prédictive d'un modèle est sa performance *hors échantillon* (parfois appelée *corrélation contre-validée*). En d'autres termes, un modèle de régression réussit *trop bien* à prévoir le résultat dans l'échantillon original : la corrélation contre-validée dans un nouvel échantillon est presque toujours inférieure à celle de l'échantillon original.

Par exemple, Dawes et Corrigan ont comparé, dans plusieurs situations, des modèles équipondérés avec des modèles de régression multiple. Une de leurs études portait sur les prédictions de la note moyenne générale (GPA) de quatre-vingt-dix étudiants en psychologie à l'Université de l'Illinois, en

utilisant dix variables liées à la réussite scolaire : notes obtenues à des tests d'aptitude, notes obtenues à l'université, évaluations de personnalité par des pairs, et autoévaluations (par exemple de l'assiduité). Dans l'échantillon initial, le modèle de régression multiple obtenait une corrélation respectable de 0,69 ; mais, hors échantillon, celle-ci tombait à 0,57 (PC = 69 %).

Dawes et Corrigan ont ensuite mesuré, dans ce nouvel échantillon, la performance du modèle équipondéré. Elle s'avère sensiblement égale à celle du modèle de régression multiple : 0,60 (PC = 70 %). Des résultats similaires ont été obtenus dans de nombreuses autres études[3]. Dans l'échantillon de départ, le modèle optimal est (par construction) supérieur au modèle équipondéré ; mais cet avantage disparaît quand on mesure la performance qui compte : la prédiction hors échantillon.

Pourquoi cet étonnant résultat ? La perte de performance quand on passe de l'échantillon original à un nouvel échantillon est d'autant plus grande que l'échantillon initial est petit, car les hasards pèsent plus lourd dans les échantillons de petite taille. Le problème que Dawes a mis en lumière, et qui explique son étonnante découverte, est que les échantillons utilisés dans la recherche en sciences sociales sont généralement trop petits, si bien que l'avantage de la pondération optimale disparaît entièrement[4]. Comme le résumait Dawes : « Rien ne sert de bâtir des modèles plus précis que nos données[5]. » Si les modèles équipondérés obtiennent de bons résultats, c'est parce qu'ils ne sont pas aussi sensibles aux accidents d'échantillonnage liés à cette imprécision.

La conclusion de Dawes a une conséquence directe qui mériterait d'être plus largement connue : on peut faire des prédictions statistiques valides, même quand on ne dispose pas de données historiques sur le résultat que l'on cherche à prédire. Il suffit pour cela d'un ensemble de prédicteurs dont on peut être sûr qu'ils sont corrélés avec le résultat.

Supposons par exemple que vous deviez, comme dans l'exemple du chapitre 9, prévoir la performance de cadres qui ont été évalués sur sept

critères. Vous tenez pour acquis que ces évaluations mesurent des qualités importantes, mais vous ne disposez pas d'un historique qui vous permettrait de calculer précisément dans quelle mesure chacune de ces qualités est associée avec les performances futures. Idéalement, vous devriez suivre dans la durée le parcours d'un large échantillon de cadres afin de mesurer leurs performances réelles, et recueillir ainsi les données nécessaires à un modèle de régression multiple. Mais vous ne pouvez pas vous permettre d'attendre plusieurs années. Alors, que faire ?

Les travaux de Dawes suggèrent une réponse. Vous pouvez simplement prendre les sept évaluations, et, en guise de prédiction, utiliser leur moyenne équipondérée (ce qui ne nécessite qu'une manipulation statistique assez simple[6]). Quelle serait la fiabilité du modèle ainsi obtenu ? Dans l'étude précitée sur la prédiction de la performance des cadres, sa corrélation avec la performance réelle est de 0,25 (PC = 58 %)[7], ce qui la rend bien supérieure au jugement clinique (r = 0,15, PC = 55 %), et sans doute assez semblable à ce qu'obtiendrait un modèle de régression dans une validation hors échantillon. Mais surtout, pour construire ce modèle, vous n'avez besoin ni de calculs complexes, ni de données inaccessibles.

Pour reprendre la formule de Dawes, devenue virale chez les chercheurs qui étudient le jugement, les modèles équipondérés sont un exemple de « beauté dans la robustesse[8] ». La dernière phrase de son article fondateur en offre un résumé lapidaire : « Toute l'astuce, c'est de décider quelles variables observer, et de savoir faire une addition[9]. »

## *Au bout de la simplicité : les modèles frugaux*

Une forme plus radicale encore de simplification consiste à utiliser des règles très simples, qu'on appelle parfois *modèles frugaux*. Il s'agit de modèles de la réalité qui semblent ridiculement simplistes, du niveau d'un calcul de coin de table. Mais dans certaines situations, ils peuvent produire des prédictions d'une justesse étonnante.

La force de ces modèles repose sur une caractéristique de la régression multiple que la plupart des gens trouvent surprenante. Supposons que vous utilisiez dans votre modèle deux variables fortement prédictives d'un résultat : leurs corrélations avec ce résultat sont respectivement de 0,60 (PC = 71 %) et 0,55 (PC = 69 %). Supposons par ailleurs que ces deux prédicteurs soient corrélés entre eux, avec une corrélation de 0,50. À votre avis, de combien pouvez-vous améliorer votre prédiction en combinant de façon optimale les deux prédicteurs ? La réponse est plutôt décevante : la corrélation sera de 0,67 (PC = 73 %), ce qui est certes mieux qu'avec le meilleur de vos deux prédicteurs de départ, mais pas beaucoup mieux.

Cet exemple illustre une règle générale : la combinaison de deux prédicteurs corrélés est à peine plus prédictive que le meilleur des deux tout seul. Or, dans la vie réelle, les prédicteurs sont presque toujours corrélés les uns avec les autres, et souvent fortement. C'est ce simple constat statistique qui plaide pour l'utilisation d'un petit nombre de prédicteurs, et donc pour les approches frugales de la prédiction. Dans certaines situations, des règles très simples, fondées sur un tout petit nombre de prédicteurs et ne nécessitant que très peu (voire pas du tout) de calculs, peuvent ainsi produire des prédictions incroyablement exactes, si on les compare à des modèles utilisant beaucoup plus de prédicteurs.

En 2020, une équipe de chercheurs a publié une étude à grande échelle sur l'application d'une approche frugale à différents problèmes de prédiction, notamment celui qui se pose à un juge quand il doit décider s'il libère un prévenu sous caution ou au contraire s'il le place en détention provisoire[10]. Si le juge refuse la liberté sous caution à un prévenu qui ne

présente pas un risque important, celui-ci sera inutilement incarcéré, ce qui est évidemment fâcheux pour lui (et coûteux pour la société). Si, à l'inverse, le juge fait confiance au mauvais prévenu, ce dernier risque de prendre la fuite avant son procès, voire de commettre un autre crime. Implicitement, la décision du juge est donc une forme de prédiction : c'est une évaluation du niveau de risque que présente chaque prévenu.

Le modèle frugal construit par les chercheurs pour estimer ce risque n'utilise que deux indicateurs, dont on sait qu'ils sont associés à la non-comparution devant un tribunal : l'âge (les prévenus plus âgés présentent un moindre risque de fuite) et le nombre de défauts de comparution dans le passé (les personnes qui ont déjà manqué à cette obligation ont tendance à récidiver). Un simple barème traduit chacun de ces deux indicateurs en un certain nombre de points, et la somme de ces points fait office d'évaluation du risque. Nul besoin d'un ordinateur, ni même d'une calculatrice.

Testé sur des données réelles, ce modèle frugal obtient d'aussi bons résultats que des modèles utilisant un nombre beaucoup plus élevé de variables pour prévoir le risque de non-comparution. Surtout, il fait mieux que la quasi-totalité des juges humains.

Le même genre de modèle frugal, n'utilisant pas plus de cinq prédicteurs et traduisant simplement les valeurs de ces prédicteurs en petits nombres entiers (entre – 3 et + 3), peut être utilisé dans des tâches de prédiction variées. Les chercheurs en ont testé l'efficacité pour évaluer un risque de crédit, pour juger de la gravité d'une tumeur à partir d'une mammographie, ou pour diagnostiquer une maladie cardiaque. Dans toutes ces tâches, le modèle frugal a obtenu des résultats aussi bons que des modèles de régression linéaire (quoique généralement en deçà de ceux qu'obtient le *machine learning*).

Une étude distincte a démontré la puissance des modèles frugaux en examinant un problème judiciaire similaire mais distinct : la prévision de la récidive[11]. En utilisant deux prédicteurs seulement pour évaluer le risque de

récidive d’un prévenu, un modèle frugal parvient à égaler le niveau de validité que revendique un outil existant, outil qui n’utilise pas moins de cent trente-sept variables[12]. Les deux prédicteurs retenus (l’âge et le nombre de condamnations antérieures) sont d’ailleurs voisins des deux facteurs utilisés dans le modèle de libération sous caution : ce sont, là aussi, ceux dont le lien statistique avec le récidivisme est le plus solidement établi[13].

Le principal attrait des modèles frugaux tient à leur simplicité et à leur transparence, qui peuvent en faciliter l’acceptation. Et ils peuvent représenter un excellent compromis, puisque le « prix à payer » pour obtenir ces avantages, c’est-à-dire le sacrifice à consentir en termes d’exactitude prédictive, est souvent bien moins élevé qu’on ne le croit.

## *Vers plus de complexité : l’apprentissage automatique*

Pour la seconde partie de cette exploration, nous allons prendre la direction opposée sur l’échelle de la complexité. Qu’arriverait-il si nous pouvions utiliser beaucoup plus de prédicteurs et réunir beaucoup plus de données sur chacun d’entre eux ? Et que se passerait-il si nous étions capables de déceler entre ces prédicteurs et le résultat cible des relations complexes – si complexes qu’aucun humain ne pourrait les détecter à l’œil nu ? Pourrions-nous alors modéliser ces relations pour faire de meilleures prédictions ? Telle est, en substance, la promesse de l’intelligence artificielle.

Il faut pour réaliser ce type d’analyse de très grandes bases de données, et la croissance explosive de la quantité de données disponibles est l’une des principales causes du progrès rapide de l’intelligence artificielle ces dernières années[14]. Avec de grandes bases de données, il devient notamment possible d’identifier automatiquement des « jambes cassées ».

Cette formule quelque peu énigmatique trouve son origine dans un cas de figure imaginé par Meehl. Prenant l'exemple d'un modèle conçu pour prédire la probabilité que des gens aillent au cinéma le soir, Meehl avait fait la remarque suivante : indépendamment de votre confiance dans le modèle, si vous apprenez qu'un individu donné vient de se casser une jambe, vous savez sans doute mieux que le modèle à quoi risque de ressembler sa soirée.

Le principe de la jambe cassée est riche d'enseignements pour les décideurs qui utilisent des modèles simples. Il leur dit en effet quand suivre la recommandation du modèle et quand passer outre. Si vous possédez des informations décisives que le modèle n'a pas pu prendre en considération, vous êtes devant un cas de jambe cassée, et vous aurez raison d'ignorer la recommandation du modèle. Si, en revanche, votre intuition vous suggère que le modèle se trompe alors que vous ne disposez pas d'informations décisives dont le modèle est privé, la situation est bien différente. Votre tentation de passer outre à la recommandation du modèle témoigne d'un pattern de jugement personnel. Or, nous l'avons vu, ce pattern personnel contient généralement plus de bruit que de subtilité, ce qui le rend très suspect. Vous devriez donc résister à la tentation de contredire le modèle : il y a de fortes chances que votre intervention dégrade la qualité de la prédiction.

Une des raisons du succès des modèles d'apprentissage automatique dans les tâches de prédiction est qu'ils sont capables de détecter ces cas de jambe cassée – et en bien plus grand nombre qu'on ne peut l'imaginer. Reprenons l'exemple d'un modèle prédisant le comportement des cinéphiles : avec une grande quantité de données sur un très grand nombre de personnes, un tel modèle pourrait apprendre, par exemple, que si une personne a été admise à l'hôpital le jour de la semaine où elle va habituellement au cinéma, il est peu probable qu'elle y aille ce soir-là. Cette manière d'améliorer la prédiction des événements rares diminue la nécessité d'un contrôle humain.

Soulignons que ce processus ne relève pas de la magie, ni même de la compréhension : il s'agit d'une simple détection de régularités statistiques. Pour impressionnant que soit l'apprentissage automatique, n'oublions pas qu'il faudra sans doute longtemps pour que l'intelligence artificielle comprenne *pourquoi* une personne qui s'est cassé une jambe n'ira pas au cinéma ce soir.

## *Un exemple : améliorer les décisions de détention provisoire*

À peu près en même temps que l'équipe de chercheurs mentionnée précédemment développait des modèles frugaux pour aider les décisions de libération sous caution, une autre équipe, conduite par Sendhil Mullainathan, formait des modèles complexes d'intelligence artificielle (IA) à la même tâche[15]. Les chercheurs avaient accès à un volume considérable de données : 758 027 décisions d'accorder ou non la liberté sous caution à un prévenu dans l'attente de son procès. Pour chaque affaire, l'équipe avait aussi accès aux informations dont disposait le juge : le délit commis par le prévenu, son casier judiciaire, et ses précédents défauts de comparution. En dehors de l'âge, aucune information démographique n'était utilisée pour « entraîner » l'algorithme (c'est-à-dire pour lui permettre « d'apprendre »).

Les chercheurs savaient aussi, pour chacune des affaires, si le prévenu avait été libéré sous caution, et, dans ce cas, s'il avait manqué à son obligation de comparaître ou fait l'objet d'une nouvelle arrestation. (Sur l'ensemble des prévenus, 74 % avaient été libérés sous caution ; parmi ceux-ci, 15 % ne s'étaient pas représentés à leur procès, et 26 % avaient été arrêtés pour une autre infraction.) Munis de ces données, les chercheurs

entraînèrent un algorithme d'apprentissage automatique et évaluèrent sa performance[16]. Comme le modèle avait été construit au moyen de l'apprentissage automatique, il n'était pas limité à des combinaisons linéaires. S'il détectait dans les données des patterns complexes, il pouvait utiliser ces subtilités pour améliorer ses prédictions.

Une précision s'impose. Décider si un prévenu doit être mis en liberté sous caution s'analyse en réalité comme la combinaison de deux jugements : d'abord, un jugement du risque de non-comparution du prévenu en question ; ensuite un jugement du seuil maximum de risque acceptable, c'est-à-dire du niveau de risque au-delà de la détention provisoire, s'impose. Le premier jugement est prédictif, mais le second est évaluatif : il s'agit d'un arbitrage entre deux maux, les non-comparutions d'une part et les détentions préventives inutiles d'autre part. Aucun modèle ne peut faire cet arbitrage. Le modèle de Mullainathan et de ses collègues était donc conçu seulement pour reproduire le premier des deux jugements : une prévision chiffrée de la probabilité de défaut de comparution.

La performance prédictive d'un tel modèle ne peut pas être comparée directement à un résultat, puisqu'on ne connaîtra jamais la probabilité exacte qu'un détenu particulier prenne la fuite. Mais il est possible d'évaluer le modèle en simulant différentes valeurs possibles du jugement évaluatif qui concerne le seuil de risque. C'est donc ce que firent les chercheurs, et leur conclusion fut claire : quel que soit le seuil de risque choisi, leur modèle prévoit mieux que les juges humains.

L'équipe de Mullainathan simula par exemple un seuil de risque tel que le nombre de prévenus envoyés en détention provisoire reste le même que celui auquel arrivaient en réalité les juges. Dans cette hypothèse, calculèrent-ils, le taux de récidive pourrait être réduit de 24 % : le nombre de personnes incarcérées serait le même, mais ce seraient majoritairement celles qui avaient le plus de probabilité de récidiver. Inversement, si le seuil de risque était fixé de manière à réduire le plus possible le nombre de

détentions provisoires sans accroître le taux de récidives et de non-comparutions, le nombre d'incarcérations pourrait être réduit de 42 %. En d'autres termes, quel que soit le jugement évaluatif retenu (jugement qui ne dépend pas du modèle), le modèle d'apprentissage automatique était bien plus performant que les juges humains pour le jugement prédictif consistant à identifier les prévenus présentant le risque le plus élevé.

Le modèle d'apprentissage automatique était également beaucoup plus performant que des modèles linéaires utilisant les mêmes informations. L'explication qu'en donnent les auteurs est riche d'enseignements : « L'algorithme d'apprentissage automatique trouve un signal significatif dans des combinaisons de variables qui pourraient autrement passer inaperçues[17]. » Cette capacité de l'algorithme à détecter des patterns qui échappent aux autres méthodes est particulièrement prononcée pour les prévenus que l'algorithme juge comme présentant un très haut risque. Il y a dans les données des patterns qui sont rares, mais caractéristiques d'un risque élevé. Ces patterns rares mais décisifs, que l'algorithme parvient seul à identifier, sont des exemples typiques de ce que nous avons appelé plus haut des « jambes cassées ».

Forts de leur algorithme d'apprentissage automatique, les chercheurs ne se sont pas contentés de bâtir un modèle de décision optimale : ils ont aussi construit, pour chaque juge humain, un « modèle du juge », analogue à ceux que nous avons décrits dans le chapitre 9, à cette importante nuance près que les modèles, cette fois, n'étaient pas limités à des combinaisons linéaires simples. En utilisant ces « modèles de juges » et en les appliquant à l'ensemble des données (alors qu'en réalité chaque juge ne traite bien sûr que d'un petit nombre d'affaires), il devient possible de simuler les décisions que les juges auraient prises s'ils avaient tous jugé toutes les affaires. On peut donc comparer leurs décisions, ce qui revient à simuler un audit de bruit.

Les résultats de cette simulation indiquent un niveau important de bruit systémique dans les décisions des juges. Une partie de ce bruit est, comme on l'imagine aisément, du bruit de niveau : certains juges ont la main plus lourde que d'autres. Ainsi, quand les juges sont classés par niveau de clémence, le quintile le plus clément (c'est-à-dire les 20 % de juges les moins enclins à infliger une détention provisoire) incarcère 17 % des prévenus ; pour le quintile le moins clément, ce chiffre est de 39 %. Cependant, indépendamment de ces écarts de clémence moyenne, les juges présentent aussi des profils de jugement très différents. Un même prévenu, jugé fiable par un juge donné, peut être considéré comme présentant un risque élevé par un autre juge, même lorsque ce second juge n'est, en moyenne et en général, pas plus sévère. En d'autres termes, les données établissent clairement la présence de bruit de pattern. Une analyse plus détaillée montre que les différences entre les affaires expliquent 67 % de la variance dans les jugements des juges : 33 % de cette variance est donc du bruit systémique. Celui-ci est constitué pour l'essentiel (79 %) de bruit de pattern, le reste étant du bruit de niveau [18].

## *Algorithmes et discriminations*

Mullainathan et ses collègues soulignent une autre conclusion importante : la performance prédictive supérieure du programme d'apprentissage automatique ne se fait pas au détriment d'autres objectifs identifiables que les juges pourraient poursuivre, et notamment de l'égalité de traitement entre groupes ethniques. Ce point est important, car, en théorie au moins, même un algorithme qui n'utilise pas de données ethniques peut aggraver par inadvertance la discrimination raciale. Des disparités de traitement pourraient survenir si le modèle utilisait des prédicteurs fortement corrélés avec l'origine ethnique (c'est le cas d'indicateurs d'apparence aussi

innocente que le code postal de résidence). Elles pourraient également se faire jour si l'algorithme était entraîné sur des données elles-mêmes biaisées. Si, par exemple, le nombre d'arrestations passées était utilisé comme prédicteur, et si l'on suppose que la fréquence de ces arrestations est elle-même affectée par la discrimination raciale, alors l'algorithme produira lui aussi de la discrimination.

Bien que l'apparition de discriminations soit donc bien un risque en principe, l'étude de Mullainathan et de ses coauteurs démontre que, dans le cas d'espèce, les décisions de l'algorithme ne sont pas plus discriminatoires que celles des juges. Au contraire, elles le sont moins. Par exemple, quand le seuil de risque est fixé de telle manière que le taux de criminalité reste constant par rapport à celui que produisent les décisions des juges, l'algorithme envoie en prison 41 % de prévenus de couleur de moins. Des résultats similaires apparaissent dans les autres scénarios. En somme, l'utilisation de l'algorithme apporte bien des gains en exactitude, mais n'exacerbe pas pour autant les disparités raciales. De surcroît, comme l'a également montré l'équipe de chercheurs, il serait facile d'utiliser l'algorithme pour réduire ces disparités.

Une autre étude, dans un domaine différent, montre que les algorithmes peuvent simultanément produire des prévisions plus exactes et réduire les discriminations. Bo Cowgill, professeur à la Columbia Business School, a travaillé sur le recrutement d'informaticiens par une grande entreprise technologique[19]. Comme beaucoup d'entreprises qui reçoivent de très nombreuses candidatures, cette entreprise doit trier les CV des candidats afin de sélectionner ceux qui seront invités à un entretien. Pour automatiser cette tâche, Cowgill a créé un algorithme d'apprentissage automatique qui passe à la loupe les CV des candidats, et l'a formé sur plus de trois cent mille CV reçus et évalués par l'entreprise. Il a ainsi montré que, par rapport aux candidats sélectionnés par des recruteurs humains, ceux sélectionnés par l'algorithme étaient de meilleurs choix : non seulement ils avaient 14 %

de chances en plus de se voir proposer un poste après l'entretien, mais ils avaient 18 % de chances en plus d'accepter cette proposition. Indépendamment de sa qualité de prévision, l'algorithme se distinguait aussi en choisissant des candidatures plus diverses en termes de race, de genre, et d'autres caractéristiques. En particulier, il avait davantage tendance à sélectionner des profils « non traditionnels » : des candidats qui n'étaient pas diplômés d'une école d'élite, qui n'avaient pas d'expérience professionnelle, ou qui n'étaient pas recommandés par un employé actuel de l'entreprise. Là où les humains avaient tendance à préférer les CV qui cochaient toutes les cases du profil « type » de l'informaticien, l'algorithme semblait donner à chaque prédicteur son juste poids, et produire ainsi non seulement plus de justesse, mais aussi plus de justice.

Pour être clair, ces exemples ne prouvent pas que les algorithmes sont *toujours* justes, non biaisés et non discriminants. Le piège type est celui de l'algorithme qui doit prédire la réussite de candidats, mais qui est entraîné à partir d'un échantillon de décisions de promotion passées : inévitablement, un tel algorithme reproduira l'ensemble des biais humains qui ont pesé sur ces décisions. De manière générale, il est possible, et peut-être trop facile, de construire un algorithme qui perpétue les disparités raciales ou de genre. Les exemples de ce type de problème sont légion ; et leur visibilité explique les inquiétudes croissantes quant à la présence de biais dans la prise de décision algorithmique. Mais des exemples alarmants ne démontrent pas une fatalité. Il existe aussi des algorithmes qui sont à la fois plus exacts que les juges humains et plus équitables.

## *Pourquoi si peu d'algorithmes ?*

Pour résumer ce tour d'horizon des techniques de prédiction « mécaniques », deux raisons distinctes expliquent la supériorité des règles

et algorithmes de toutes sortes par rapport au jugement humain.

Premièrement, comme nous l'avons vu au chapitre précédent, toutes les techniques de prédiction mécanique, et pas seulement les plus récentes et les plus sophistiquées, ont un avantage de taille par rapport au jugement humain : l'absence de bruit. Les patterns de jugement personnels et le bruit occasionnel pèsent lourd sur la qualité du jugement, si lourd que la simplicité et l'absence de bruit se traduisent par des gains importants. Dès lors qu'elles sont appliquées avec constance – c'est-à-dire sans bruit –, il suffit donc souvent que des règles soient simplement sensées pour faire mieux que le jugement humain.

Deuxièmement, les données sont parfois suffisamment riches pour que des techniques d'intelligence artificielle sophistiquées y détectent des patterns valides, et surpassent ainsi la force prédictive des modèles simples. Dans ce cas, l'avantage de ces modèles sur le jugement humain n'est pas seulement l'absence de bruit, mais aussi la capacité d'exploiter beaucoup plus d'informations.

Compte tenu de ces avantages et du nombre d'études qui en attestent, une question s'impose : pourquoi les algorithmes ne sont-ils pas plus largement utilisés pour tous les types de jugements professionnels qui sont examinés dans ce livre ? En effet, malgré les vifs débats de principe suscités par les algorithmes, leur utilisation – en dehors de quelques domaines d'application bien particuliers – reste limitée. Beaucoup d'experts choisissent d'ignorer le résultat du match opposant jugement clinique et prédiction mécanique : ils se fient à leur intuition et ne croient pas un instant que des machines puissent faire mieux qu'eux. Pour eux, utiliser des algorithmes serait un choix déshumanisant, voire une abdication de leur responsabilité.

Dans le diagnostic médical, par exemple, l'utilisation d'algorithmes est loin d'être habituelle, malgré des avancées impressionnantes. Peu d'organisations utilisent les algorithmes dans leurs décisions de recrutement

et de promotion de leur personnel. Les patrons des studios de Hollywood choisissent les films qu'ils produiront sur la base de leur jugement et de leur expérience, pas en fonction d'une formule. Les maisons d'édition font de même. Et si l'histoire d'une équipe de base-ball obsédée de statistiques (les Athletics d'Oakland), racontée par Michael Lewis dans *Moneyball*, a fait si forte impression, c'est précisément parce que la rigueur algorithmique que cette équipe a adoptée demeure l'exception, et non la règle, dans le processus de décision des clubs sportifs. Entraîneurs, sélectionneurs et managers se fient le plus souvent à « leurs tripes », certains que l'analyse statistique ne pourra jamais remplacer leur jugement.

Dans un article publié en 1996, Meehl et un coauteur énuméraient, pour les réfuter, pas moins de dix-sept objections que les professionnels – psychiatres, médecins, juges et autres – formulent à l'encontre de la prédiction mécanique[20]. Les auteurs concluaient que ces résistances pouvaient s'expliquer par plusieurs facteurs psychologiques, dont la « peur d'être rendu inutile par la technologie », une « formation insuffisante » et une « antipathie générale pour l'informatique ».

Depuis lors, les chercheurs ont identifié bien d'autres sources de résistance. Nous n'ambitionnons pas d'en donner ici un panorama complet. Notre but dans ce livre est de faire des propositions pour améliorer le jugement des humains, pas de prôner « le remplacement des individus par des machines », comme l'aurait dit le juge Frankel.

Toutefois, ces analyses de la résistance des humains à la prédiction mécanique sont intéressantes par ce qu'elles nous apprennent plus généralement sur le jugement humain. De récentes recherches ont notamment mis en lumière un point essentiel : les gens ne se méfient pas *systématiquement* des algorithmes. Quand on leur donne le choix entre prendre conseil auprès d'un humain et auprès d'un algorithme, par exemple, ils préfèrent souvent l'algorithme[21]. L'*aversion aux algorithmes* ne s'exprime pas toujours par le refus catégorique d'adopter de nouveaux

outils d’aide à la décision. En réalité, les gens sont souvent prêts à donner sa chance à l’algorithme. Mais ils ont tendance à lui retirer leur confiance dès qu’ils constatent qu’il fait des erreurs[22].

D’une certaine manière, cette réaction paraît tout à fait sensée : pourquoi perdre son temps avec un algorithme auquel on ne peut pas faire confiance ? Nous sommes tout à fait conscients que l’erreur est humaine, si conscients, même, que nous entendons en conserver le monopole. Nous attendons des machines qu’elles soient parfaites ; et si cette attente est déçue, nous n’en voulons plus[23].

Le problème que pose cette attente de perfection, c’est qu’elle conduit à de mauvais choix. Inévitablement, elle nous conduit à nous méfier des algorithmes et à continuer de faire confiance à notre jugement, même quand il produit manifestement de moins bons résultats. Cette attitude est profondément ancrée : il est peu probable que nous la surmontions tant que les algorithmes, dans un domaine donné, n’atteindront pas une exactitude prédictive proche de la perfection.

---

Les règles et les algorithmes sont souvent supérieurs au jugement humain, mais ils ne le remplaceront bien sûr jamais entièrement. Fort heureusement, même quand nous n’adoptons pas les règles et les algorithmes, nous pouvons nous en inspirer. Une bonne partie de ce qui les rend supérieurs à notre intuition peut être reproduit dans nos méthodes de jugement. Nous n’utiliserons jamais autant d’informations qu’un modèle d’intelligence artificielle, mais nous pouvons essayer d’imiter la simplicité et l’absence de bruit des modèles frugaux. Et si nous parvenons à adopter des méthodes qui réduisent le bruit, nos jugements prédictifs devraient s’améliorer. Ce sera le sujet principal de la partie V.

## À propos des règles et des algorithmes

« Les algorithmes d'apprentissage automatique font mieux que les humains et que les modèles simples. Mais il n'est pas forcément nécessaire d'utiliser des masses de données : même les règles les plus simples ont sur les juges humains l'avantage de n'être pas sujettes au bruit. »

« Les règles ont un autre atout : elles n'essaient pas de suivre des intuitions complexes, séduisantes, mais généralement invalides. Face à un problème donné, aurions-nous intérêt à utiliser une règle, plutôt qu'à juger au cas par cas ? »

« Comme nous manquons de données sur le résultat que nous voulons prédire, pourquoi ne pas utiliser un modèle équipondéré ? Il ne fera peut-être pas aussi bien qu'un modèle optimal, mais il fera toujours mieux que le jugement humain au cas par cas. »

« Vous n'êtes pas d'accord avec la prévision du modèle. D'accord. Mais est-ce parce que vous avez trouvé une "jambe cassée", ou seulement parce que la prévision ne vous plaît pas ? »

« L'algorithme fait des erreurs, bien sûr. Mais si les juges humains en font encore plus, à qui faut-il se fier ? »

CHAPITRE 11

# L’ignorance objective

Il nous est souvent arrivé de présenter à un auditoire de dirigeants le contenu des deux derniers chapitres, et en particulier les résultats que nous avons exposés sur les performances limitées du jugement humain. La plupart de ces travaux remontent déjà à un demi-siècle, et on peut donc penser que la plupart des décideurs en ont au moins entendu parler. Manifestement, pourtant, beaucoup sont parvenus à n’en tenir aucun compte.

Certains de ces cadres et de ces dirigeants d’entreprise nous ont expliqué, non sans une certaine fierté, qu’ils jugent indispensable de se fier à leur instinct, à leurs « tripes », plutôt qu’à des analyses en tous genres. Beaucoup d’autres partagent cet avis, même s’ils l’expriment de manière moins catégorique. La recherche en matière de décision en entreprise montre que les cadres et les dirigeants, en particulier les plus expérimentés et ceux qui occupent les postes les plus élevés, font largement appel à quelque chose qu’ils appellent leur intuition, leur flair, leur discernement, ou simplement leur jugement (terme qu’ils utilisent alors dans une acception très différente de celle de ce livre[1]).

En somme, les décideurs se plaisent à écouter leur intuition, et la plupart ont l’air très satisfaits de ce qu’ils entendent. Ce qui soulève une question :

quand ces décideurs si sûrs d'eux tendent ainsi l'oreille, qu'est-ce qu'ils entendent, exactement ?

Une étude sur l'intuition dans la prise de décision managériale la définit comme suit : « Jugement relatif à un plan d'action, qui vient à l'esprit avec une aura ou une conviction de justesse ou de plausibilité, mais sans raisons ni justifications clairement exprimées : fondamentalement, on "sait" mais sans savoir pourquoi[2]. » Nous pensons que ce sentiment de « savoir sans savoir pourquoi » est en réalité le *signal intérieur* de fin de jugement que nous avons évoqué au chapitre 4.

Le signal intérieur est une récompense autoadministrée, et ceux qui formulent un jugement travaillent dur (ou, parfois, pas assez dur) pour l'obtenir. C'est une forme de satisfaction, un sentiment agréable de cohérence, dans lequel les éléments considérés et le jugement rendu « sonnent bien ». Toutes les pièces du puzzle ont l'air de s'assembler à la perfection. (Nous verrons que ce sentiment de cohérence vient souvent du fait qu'on dissimule ou qu'on ignore des éléments qui n'y trouveraient pas leur place.)

La valeur du signal intérieur, mais aussi son danger, tiennent à la manière dont il est interprété. Celui qui ressent ce signal ne l'interprète pas comme un sentiment mais comme une conviction. Le signal intérieur est un sentiment (« ça semble bon »), mais ce sentiment se fait passer pour une conclusion rationnellement atteinte qui justifierait un jugement confiant (« Je sais, même si je ne sais pas pourquoi »).

Hélas, la confiance n'est pas une garantie d'exactitude, et nombreuses sont les prévisions confiantes qui se révèlent fausses. Non seulement le biais et le bruit nous font commettre des erreurs, mais il existe une limite absolue à notre capacité de prédiction. Cette limite, que nous appelons l'*ignorance objective*, est le sujet de ce chapitre.

# *L'ignorance objective*

Voici un problème que vous pourriez rencontrer si vous produisez des jugements prédictifs répétés. Ces jugements peuvent être de n'importe quel type – par exemple, choisir un portefeuille d'actions ou prédire les performances de sportifs professionnels. Mais par souci de simplicité, nous allons reprendre l'exemple déjà utilisé dans le chapitre 9 : la sélection de candidats à l'embauche. Imaginez que vous ayez évalué, au fil des années, une centaine de candidats ; et que vous ayez maintenant la possibilité d'évaluer la qualité de vos décisions de recrutement en rapprochant les prédictions que vous aviez faites des évaluations de performance qu'ont obtenues les candidats depuis leur embauche. La hiérarchie que votre jugement prédictif établit entre les candidats correspondra-t-elle à celle de leurs évaluations ? Plus précisément, chaque fois que vous comparez deux candidats choisis au hasard parmi la centaine que vous avez recrutés, quelle est la probabilité que celui dont vous pensiez qu'il avait le meilleur potentiel soit effectivement celui qui a le mieux réussi dans son poste ?

Nous avons souvent réalisé des sondages informels sur cette question auprès de groupes de cadres et de dirigeants d'entreprise. Les réponses les plus fréquentes se situent aux alentours de 75-85 %. Dans nos conversations en tête à tête, où nos interlocuteurs ne craignent pas de paraître immodestes aux yeux de leurs collègues, le sentiment de confiance est souvent plus élevé encore.

Comme vous avez reconnu dans cette question ce que nous avons appelé le pourcentage de concordance, vous comprendrez aisément le problème que pose cette autoévaluation. Un PC de 80 % correspond approximativement à une corrélation de 0,80. Il est rarissime de rencontrer une telle puissance prédictive dans le monde réel. En matière de sélection de personnel, une récente méta-analyse compilant des dizaines d'études montre que la performance des juges humains est très loin de ce chiffre : en

moyenne, la corrélation entre leurs prédictions et les résultats est de 0,28 (PC = 59 %)[3].

À première vue, ce chiffre peut sembler décevant. Il suffit pourtant de réfléchir un instant au défi que représente la sélection à l'embauche pour se rendre compte qu'il n'a rien d'étonnant. Une personne qui débute aujourd'hui dans un nouvel emploi va se heurter à de nombreuses difficultés et rencontrer de nombreuses opportunités, et l'intervention du hasard peut affecter les unes comme les autres. La nouvelle recrue peut tomber sur une supérieure hiérarchique qui croira en elle, qui saura lui donner sa chance, qui encouragera son travail et contribuera à conforter sa confiance en soi et sa motivation. Elle peut aussi avoir moins de chance et, sans que ce soit sa faute, commencer sa carrière par un échec démoralisant. Dans sa vie personnelle, également, peuvent se produire des événements qui affecteront sa performance au travail. Aucune de ces circonstances et aucun de ces événements ne peuvent être prévus aujourd'hui : ni par vous, ni par personne, ni par le meilleur modèle prédictif au monde. Cette incertitude irréductible comprend tout ce qui ne peut pas être connu à un moment donné, mais qui affecte néanmoins le résultat que nous essayons de prévoir.

De surcroît, il existe beaucoup d'informations sur un candidat qui pourraient en principe être connues, mais qui ne le sont pas du recruteur au moment où il produit son jugement. Peu importe que ces lacunes soient dues au manque d'outils de sélection suffisamment prédictifs, à une décision rationnelle de ne pas dépenser plus pour obtenir davantage d'informations, ou à une négligence coupable dans la recherche de renseignements. Quelle qu'en soit la raison, le décideur est dans une situation d'information imparfaite.

L'incertitude irréductible (ce qu'il est impossible de savoir) et l'information imparfaite (ce qui pourrait être connu mais ne l'est pas) se conjuguent pour rendre illusoire l'idée d'une prédiction parfaite. Ces limites

sont indépendantes du jugement. Elles n'ont rien à voir avec le biais et le bruit qui l'affectent. Ce sont des caractéristiques de la situation, de la tâche de prédiction en tant que telle. L'ignorance objective d'éléments importants limite notre capacité à prévoir l'avenir avec exactitude. (Notons que nous utilisons le mot *ignorance* là où l'on utilise communément celui d'*incertitude* : cette liberté terminologique limite le risque de confusion avec le bruit, qui est lui aussi une forme d'incertitude, portant non sur les états du monde, mais sur les jugements.)

Il y a davantage d'informations (et donc moins d'ignorance objective) dans certaines situations que dans d'autres. Dans la majorité des métiers, la plupart des jugements professionnels sont corrects. Les médecins diagnostiquent sans difficulté un grand nombre de maladies. Et dans un grand nombre d'affaires, les avocats sont capables de prévoir assez précisément ce que vont décider les juges. De façon générale, cependant, on peut s'attendre à ce que les gens qui formulent des prévisions sous-estiment leur ignorance objective. L'excès de confiance est un des biais cognitifs les mieux documentés. En particulier, le jugement que l'on porte sur sa propre capacité à faire des prédictions précises, même à partir d'informations limitées, pèche notoirement par excès de confiance[4]. Ce que nous avons dit du bruit en matière de jugement peut se dire aussi de l'ignorance objective : *dès lors qu'il y a prévision, il y a ignorance, et plus qu'on ne l'imagine.*

## *Des experts trop sûrs d'eux*

Un de nos bons amis, le psychologue Philip Tetlock, est armé tout à la fois d'une passion pour la quête de vérité et d'un sens de l'humour malicieux. En 2005, il a publié un livre intitulé *Expert Political Judgment* [« Le Jugement politique des experts »]. Malgré l'apparente neutralité de ce titre,

l’ouvrage dressait un réquisitoire impitoyable contre les experts et leurs prédictions concernant les événements politiques.

L’analyse de Tetlock portait sur les prévisions de près de trois cents experts : de grands journalistes, des universitaires respectés, des conseillers politiques introduits au plus haut niveau de l’État. Tetlock s’est posé une question simple : leurs prévisions politiques, économiques et sociales étaient-elles justes ? Et comme ce sont souvent des prévisions à long terme, Tetlock s’arma de patience : son étude couvre deux décennies.

La principale conclusion de Tetlock est que les prévisions faites par les prétendus experts à propos des grands événements politiques étaient d’une médiocrité insigne. Une des phrases du livre est devenue célèbre aux États-Unis : « L’expert moyen est à peu près aussi précis qu’un chimpanzé qui joue aux fléchettes. » De façon plus précise, le message était que les experts qui gagnent leur vie « en commentant ou en proposant des conseils sur les grandes tendances économiques et politiques » ne déchiffraient pas mieux « les situations nouvelles que les journalistes ou les lecteurs attentifs du *New York Times*[5] ». Bien sûr, les experts savaient raconter de belles histoires. Ils étaient capables d’analyser une situation, de brosser un tableau convaincant de la manière dont elle allait évoluer, et de réfuter avec beaucoup d’assurance les objections de ceux qui n’étaient pas d’accord avec eux sur les plateaux de télévision. Mais savaient-ils vraiment ce qui allait se passer ? Non.

Pour parvenir à cette conclusion, Tetlock dut, justement, couper court au *storytelling* des experts. Sur chaque question, il leur demandait d’assigner une probabilité à trois résultats possibles : le *statu quo*, plus de quelque chose ou moins de quelque chose. Un chimpanzé lançant des fléchettes « choisirait » chacun de ces résultats avec la même probabilité – un tiers – indépendamment de la réalité. Les experts de Tetlock faisaient à peine mieux : en moyenne, ils attribuaient des probabilités légèrement supérieures aux événements qui allaient se produire qu’à ceux qui ne se produiraient

pas. Mais la caractéristique la plus marquante de leur travail était leur confiance excessive dans la précision avec laquelle ils pouvaient formuler des prédictions. D'ailleurs, parmi les experts, ceux qui avaient des idées arrêtées et des théories claires sur la manière dont tourne le monde étaient ceux qui avaient le plus confiance en eux – et ceux qui se trompaient le plus.

Ce que les travaux de Tetlock nous enseignent, c'est qu'il est tout simplement impossible de faire des prévisions précises à long terme sur des événements particuliers. Le monde est trop chaotique pour cela. Des événements imprévisibles et mineurs peuvent avoir des conséquences trop considérables. Songez par exemple à n'importe quel grand personnage de l'histoire (ou d'ailleurs à toute autre personne) : à l'instant de sa conception, il y avait une chance égale pour que ce personnage naisse du sexe opposé. Quand vous considérez l'avenir, des événements imprévisibles vont immanquablement se produire, et les conséquences de ces événements imprévisibles sont elles aussi imprévisibles. De ce fait, plus l'on regarde loin dans le futur, plus l'ignorance objective s'accumule.

La limite du jugement politique des experts ne réside donc pas tant dans les limites cognitives des prévisionnistes que dans leur irrémédiable ignorance objective de l'avenir. C'est pourquoi on ne devrait pas reprocher aux experts de se tromper dans leurs prévisions à long terme. Ce dont on devrait leur tenir rigueur, ce n'est pas de ne pas réussir : c'est d'essayer. L'erreur du prévisionniste, c'est de tenter l'impossible et de croire qu'il peut y arriver.

Quelques années après avoir ainsi démontré qu'une grande partie des tentatives pour formuler des prévisions à long terme sont vaines, Tetlock fit équipe avec son épouse Barbara Mellers pour étudier des prévisions à relativement court terme (généralement moins d'un an). En étudiant différents grands événements mondiaux, les deux chercheurs découvrirent que de telles prévisions à court terme sont difficiles, mais pas impossibles.

Certains individus, auxquels ils donnèrent le nom de *superprévisionnistes*, s'y distinguent, et font parfois même mieux que les professionnels des services de renseignement. Ces résultats confortent l'idée que l'ignorance objective s'accroît à mesure que l'horizon de prévision s'éloigne. Nous reviendrons sur ces superprévisionnistes dans le chapitre 21.

## *Des humains médiocres, et des modèles à peine meilleurs*

Les premiers travaux de Tetlock démontraient l'incapacité générale des experts à obtenir de bons résultats en matière de prévisions politiques à long terme. Il lui aurait suffi de trouver une seule personne capable de voir juste pour changer entièrement ses conclusions. Une tâche ne peut être considérée comme impossible qu'après qu'un grand nombre d'acteurs crédibles s'y sont essayés sans succès.

Par ailleurs, nous l'avons vu, le traitement automatique de l'information est souvent supérieur au jugement humain. Pour mesurer l'ignorance objective, c'est-à-dire le caractère intrinsèquement imprévisible (ou prévisible) d'un résultat, nous devons donc nous tourner vers les outils qui ont les meilleures chances de prévoir ces résultats : les règles et les algorithmes. Si ce sont les algorithmes qui font les meilleures prévisions, ce sont eux qui nous rapprochent de la limite ultime à la prévisibilité.

Les précédents chapitres vous ont peut-être donné l'impression que les algorithmes sont immensément supérieurs aux jugements prédictifs. Cette impression est trompeuse. Les modèles font toujours mieux que les humains, mais pas *beaucoup* mieux. Il n'existe pratiquement pas de situations dans lesquelles, avec les mêmes informations, les humains soient très mauvais et les modèles très bons.

Dans le chapitre 9, nous avons parlé d'une synthèse de cent trente-six études démontrant la supériorité de la prédiction mécanique sur le jugement clinique[6]. Si les preuves de cette supériorité sont en effet « massives et concordantes », l'écart de performance n'est pas très important. Dans quatre-vingt-treize de ces études, il était question de décisions binaires où l'on mesure le « taux de réussite » des cliniciens et celui des formules. Résultat : dans l'étude médiane, les cliniciens avaient raison 68 % du temps et les formules 73 %. Un autre sous-ensemble de trente-cinq études utilisait le coefficient de corrélation pour mesurer la justesse des prédictions : les cliniciens obtenaient 0,32 (PC = 60 %) et les formules 0,56 (PC = 69 %). On le voit, sur un critère comme sur l'autre, les formules font mieux que les cliniciens, parfois bien mieux, mais il est frappant de constater que même les prévisions mécaniques ont une validité limitée. Les modèles volent un peu plus haut que les humains, mais le plafond de la prévisibilité reste assez bas.

Qu'en est-il de l'intelligence artificielle ? Comme nous l'avons noté, l'IA fait généralement mieux que les modèles plus simples. Mais dans la plupart des applications prédictives, ses performances restent elles aussi très loin de la perfection. Prenons par exemple l'algorithme de prévision des décisions de libération sous caution dont nous avons parlé dans le chapitre 10. Nous avons vu qu'à nombre constant de prévenus placés en détention provisoire, l'algorithme pouvait réduire le taux de criminalité de 24 %. C'est certes une amélioration considérable par rapport aux juges humains, mais si l'algorithme avait pu prédire avec une justesse parfaite quels prévenus récidiveraient, il aurait pu réduire le nombre de crimes bien davantage – voire, en principe, le ramener à zéro. Ce n'est pas pour rien que *Minority Report*, où l'on prévoit les crimes avant qu'ils ne soient commis, est de la science-fiction : dans la prédiction des comportements humains, il y a une part considérable d'ignorance objective.

Sendhil Mullainathan, dont nous avons déjà mentionné les travaux, a réalisé, en collaboration avec Ziad Obermeyer, une étonnante étude sur le diagnostic des crises cardiaques[7]. Quand des patients présentent des signes de possible crise cardiaque, les médecins urgentistes doivent décider de prescrire ou non des examens supplémentaires. En principe, les patients ne devraient être examinés que si le risque de crise cardiaque est suffisamment élevé : l'examen est non seulement coûteux, mais aussi invasif et risqué, et il vaut donc mieux en dispenser les patients présentant un risque faible. La décision d'un médecin de prescrire ou non des examens constitue donc une évaluation implicite de la probabilité que les symptômes observés soient ceux d'une crise cardiaque.

Les chercheurs ont élaboré un modèle d'intelligence artificielle pour faire cette évaluation. Ce modèle utilise plus de 2 400 variables et est basé sur un large échantillon de cas (4,4 millions de consultations, pour 1,6 million de patients). Avec une telle quantité de données, on peut penser que le modèle s'approche du plafond de prévisibilité, c'est-à-dire de la limite qu'impose l'ignorance objective.

Le modèle d'intelligence artificielle s'avère nettement plus exact que les médecins, ce qui, à la lumière des résultats que nous avons déjà passés en revue, n'est guère surprenant. Mais que vaut sa prévision dans l'absolu ? Pour s'en faire une idée, on peut par exemple se pencher sur les patients que le modèle place dans le décile de risque le plus élevé. Parmi ceux-ci, 30 % souffrent effectivement d'une crise cardiaque, alors que ce chiffre n'est que de 9,3 % pour les patients situés dans les déciles médians. Cette qualité de « jugement » est impressionnante, mais très loin d'être parfaite : un modèle « omniscient » réaliserait une séparation claire et nette entre les patients qui ont une crise cardiaque et ceux qui souffrent d'une autre pathologie. On peut donc raisonnablement conclure de cette étude que les imperfections du diagnostic des médecins tiennent au moins autant à l'ignorance objective qu'aux limites de leur jugement.

## *Le déni de l'ignorance*

En insistant sur l'impossibilité de la prédiction parfaite, peut-être donnons-nous l'impression d'enfoncer une porte ouverte. Dire que l'avenir est imprévisible n'est certes pas une percée conceptuelle. Mais l'évidence de cette observation n'a d'égale que la constance avec laquelle elle est ignorée. En témoignent de très nombreuses études sur l'excès de confiance dans les prédictions.

Souvenez-vous du sondage informel que nous évoquions au début de ce chapitre, et où nous demandions à des dirigeants d'autoévaluer leurs décisions de recrutement. L'excès de confiance y joue un rôle important : nous avons vu que les gens prennent souvent le sentiment subjectif de confiance qu'ils ressentent pour un signe objectif de validité prédictive. À la lecture des profils de Nathalie et de Monica, par exemple, le signal intérieur que vous avez ressenti quand vous êtes arrivé à un jugement cohérent vous a donné confiance dans votre prévision que la performance de Nathalie serait meilleure. Mais si vous vous êtes fié à cette prédiction, vous avez été victime de l'illusion de validité : le niveau de justesse que vous pouvez obtenir avec les informations qui vous étaient données est assez faible.

Les gens qui se croient capables d'un niveau de justesse prédictive qu'il est en réalité impossible d'atteindre ne pèchent pas seulement par excès de confiance. Ils ne négligent pas seulement le risque de biais et de bruit de leurs jugements. Ils ne se jugent pas seulement supérieurs aux autres mortels. Ils croient aussi à la prévisibilité d'événements qui sont en fait imprévisibles, niant implicitement la réalité de l'incertitude. Dans les termes que nous avons utilisés ici, leur erreur est le *déni de l'ignorance*.

Le déni de l'ignorance apporte une réponse supplémentaire à la question qui intriguait tant Meehl et ses adeptes : pourquoi son message est-il resté largement ignoré, et pourquoi les décideurs continuent-ils de se fier à leur

intuition ? La réponse est liée au signal intérieur que perçoivent les décideurs quand ils écoutent leur intuition, et à la satisfaction que ce signal leur procure. Le signal intérieur est la voix de la confiance, celle du « je sais sans savoir pourquoi ». Une évaluation objective de la réelle valeur prédictive des éléments disponibles ne justifierait que rarement un tel niveau de confiance.

Renoncer à la récompense émotionnelle de la certitude intuitive n'est pas chose facile. Une remarque fréquente chez les décideurs est, à cet égard, très révélatrice : c'est dans les situations les plus incertaines, expliquent-ils souvent, qu'ils se fient le plus à leur intuition[8]. En d'autres termes, quand les faits leur refusent le sentiment de cohérence (et donc de confiance) dont ils ont tant besoin, ils se tournent vers leur intuition pour le retrouver. Le déni de l'ignorance est d'autant plus tentant que l'ignorance est grande.

Le déni de l'ignorance contribue à expliquer une autre énigme. Confrontés aux éléments que nous avons présentés ici, un grand nombre de dirigeants en tirent une conclusion paradoxale. Leurs décisions basées sur l'intuition ne sont peut-être pas parfaites, disent-ils, mais si les alternatives plus systématiques sont loin de l'être également, pourquoi les adopter ? Rappelez-vous, par exemple, que dans le domaine du recrutement, la corrélation moyenne entre les évaluations des juges humains et les performances des employés est de 0,28 (PC = 59 %). D'après la même étude, et de manière analogue aux autres exemples que nous avons mentionnés, la prédiction mécanique fait mieux, mais pas *beaucoup* mieux : sa justesse prédictive est de 0,44 (PC = 65 %). Pourquoi, demandent en substance les dirigeants, se compliquer la vie pour si peu ?

La réponse, bien sûr, c'est que s'agissant de quelque chose d'aussi important que des décisions de recrutement, cette validité supplémentaire a beaucoup de valeur. Dans d'autres domaines, les mêmes dirigeants n'hésitent pas à modifier en profondeur leur manière de travailler pour des enjeux bien moindres. Rationnellement, tout dirigeant sait que le succès

n’est jamais garanti, et que ce qu’il cherche à obtenir quand il prend une décision, c’est la meilleure probabilité possible de réussite. Entre deux billets de loterie offrant l’un 59 % de chances de gagner et l’autre 65 %, personne ne choisirait d’acheter le premier, si les deux billets étaient offerts au même prix.

La difficulté est justement que le « prix », dans cette situation, n’est pas le même. Le jugement intuitif produit sa propre récompense : le signal intérieur. Nous sommes tous prêts à faire confiance à un algorithme qui garantit un niveau très élevé d’exactitude prédictive, parce qu’un tel algorithme nous donne un sentiment de certitude qui égale ou dépasse celui produit par le signal intérieur[9]. Mais quand l’alternative est une méthode de prédiction mécanique qui ne peut pas se prévaloir d’une validité très élevée, renoncer à la récompense émotionnelle du signal intérieur est un prix élevé à payer.

---

Plus encore que les faiblesses de notre jugement, l’ignorance objective limite la qualité de nos prédictions. Mais nous sous-estimons systématiquement l’ignorance objective. C’est ce déni de l’ignorance qui nous permet de formuler des jugements même quand l’avenir est très incertain, et de ressentir alors le signal intérieur qui nous donne confiance dans nos jugements.

Cette observation a une implication pratique importante. Malgré tous les éléments qui plaident en faveur des méthodes de prédiction mécanique et algorithmique, et malgré le calcul rationnel qui montre clairement l’intérêt des améliorations incrémentales en matière de justesse prédictive, il est raisonnable de penser qu’un grand nombre de décideurs continueront à rejeter tout mode de prise de décision qui les prive de la capacité d’exercer leur intuition. Tant que les algorithmes ne sont pas proches de la perfection – et, dans de nombreux domaines, l’ignorance objective fait qu’ils ne le

seront jamais –, le jugement humain ne sera pas remplacé. C’est pourquoi il faut l’améliorer.

## À propos de l’ignorance objective

« Dès lors qu’il y a prévision, il y a ignorance, et plus qu’on ne l’imagine. »

« Avons-nous vérifié que les experts en qui nous plaçons notre confiance sont plus précis que des chimpanzés qui jouent aux fléchettes ? »

« Quand vous faites confiance à votre intuition et au signal intérieur qu’elle vous procure, votre confiance ne repose pas sur quelque chose que vous savez vraiment. Vous êtes dans le déni de votre ignorance objective. »

« Les modèles font mieux que les humains, mais pas de beaucoup. De façon générale, les jugements humains sont médiocres, et les modèles un peu meilleurs. Mais même un peu mieux, c’est mieux ! »

« Peut-être rechignerons-nous toujours à utiliser un modèle pour prendre ces décisions – nous avons trop besoin du signal intérieur pour avoir suffisamment confiance dans notre décision. Alors, assurons-nous de disposer du meilleur processus de décision possible. »

CHAPITRE 12

# Dans la vallée des Choses Normales

Les conclusions quelque peu pessimistes que nous venons de résumer soulèvent une question plus vaste : comment supportons-nous d'évoluer dans un monde où, si certains problèmes sont faciles à résoudre, beaucoup sont dominés par l'ignorance objective ? Après tout, lorsque l'étendue de l'ignorance objective est importante, nous devrions finir par nous rendre compte de la vanité de nos efforts de prévision. Or ce n'est généralement pas l'expérience que nous avons du monde. Au contraire, comme le suggérait le chapitre précédent, nous faisons montre d'une propension effrénée à formuler des prédictions hardies à partir de bien peu d'informations valables. L'échec récurrent de nos efforts pour prévoir les événements n'entame en rien notre conviction que nous les comprenons.

Dans ce chapitre, nous allons examiner ce sentiment aussi puissant que trompeur : l'impression de *comprendre* des événements que nous étions incapables de prédire. Nous l'aborderons dans deux contextes bien différents : les sciences sociales et les événements de la vie quotidienne.

## *Prédire les trajectoires de vie*

En 2020, un groupe de cent douze chercheurs, emmenés par Sara McLanahan et Matthew Salganik, professeurs de sociologie à l'Université de Princeton, publiait un article inhabituel dans les *Proceedings of the National Academy of Sciences*[1]. Leur projet consistait à déterminer dans quelle mesure les chercheurs en sciences sociales étaient capables de comprendre les trajectoires de vie des familles socialement fragiles. Sachant ce qu'ils savaient, pouvaient-ils prévoir des événements importants dans la vie d'une famille – par exemple, le succès des études secondaires de l'un des enfants, ou l'expulsion de la famille de son domicile ? Plus précisément, quel niveau de justesse prédictive les meilleurs experts pourraient-ils atteindre dans la prévision de tels événements, en utilisant toutes les informations que les sociologues collectent habituellement dans le cadre de leurs travaux ? Dans les termes que nous utilisons ici, l'objectif de l'étude était donc de mesurer le niveau d'ignorance objective qui demeure quant à ces événements de la vie, une fois que les sociologues ont fait au mieux leur travail.

Les auteurs ont exploité les résultats d'une grande étude longitudinale sur des enfants suivis de la naissance à l'âge de 15 ans, la *Fragile Families and Child Wellbeing Study* (« Étude sur les familles fragiles et le bien-être des enfants »). Cette énorme banque de données porte sur près de cinq mille enfants, pour la plupart issus de couples non mariés vivant dans de grandes villes des États-Unis, et contient plusieurs milliers d'éléments d'information sur chaque enfant et sa famille. Les données couvrent des sujets comme le niveau d'études et l'emploi des grands-parents et des parents, ou encore l'état de santé des membres de la famille, ainsi que les réponses à de multiples questionnaires et à des tests d'aptitude et de personnalité. C'est une véritable mine d'informations, dont les sociologues ont fait bon usage : plus de sept cent cinquante articles scientifiques ont été rédigés à partir des données de l'étude *Fragile Families*. Beaucoup de ces articles s'attachent à expliquer, à partir des données sur le parcours des enfants et de leurs

familles, des « résultats » importants dans la vie des enfants, comme les notes obtenues au lycée ou l'apparition d'un casier judiciaire.

L'étude dirigée par l'équipe de Princeton s'intéressait à la prévisibilité de six de ces « résultats », observés quand l'enfant était âgé de 15 ans, par exemple le fait que la famille ait été expulsée de son domicile, la moyenne générale de l'enfant au lycée, ou encore la situation matérielle des familles de manière générale. Ce qui fait son originalité, c'est sa méthode. Elle consiste à inviter des équipes de chercheurs à participer à une sorte de concours de pronostics : le but du concours est de produire des prévisions sur les six résultats retenus, en exploitant la masse de données disponibles sur chaque famille dans l'étude *Fragile Families*.

Si ce genre de concours est nouveau en sciences sociales, il est courant en informatique, par exemple sur des tâches de traduction ou de détection d'une image dans un ensemble de photographies. Dans ces compétitions, le résultat obtenu par l'équipe gagnante définit à un instant donné l'état de l'art, qui dépasse généralement celui de la compétition précédente. Dans une tâche de prédiction en sciences sociales, en revanche, on ne s'attend pas *a priori* à des améliorations rapides. On peut donc raisonnablement considérer la meilleure prédiction obtenue lors de la compétition comme une mesure de la prédictibilité du résultat à partir des données disponibles : autrement dit, comme une indication du niveau résiduel d'ignorance objective.

Le concours suscita un intérêt considérable chez les chercheurs. Le rapport final comparait les résultats de cent soixante équipes (celles sélectionnées pour concourir, parmi un nombre encore plus grand d'équipes candidates). Tous les concurrents sélectionnés étaient hautement qualifiés, et la plupart se présentaient comme des *data scientists* utilisant des techniques d'apprentissage automatique.

Le déroulement du concours comprenait deux étapes. Dans un premier temps, les équipes participantes avaient accès à toutes les données, y

compris les six résultats à prévoir, pour un sous-ensemble de familles représentant la moitié de l'échantillon total[2]. Ces « données d'apprentissage » leur permettaient d'entraîner leurs algorithmes prédictifs. Dans un second temps, les algorithmes ainsi développés étaient appliqués « hors échantillon », c'est-à-dire à un sous-ensemble différent de familles. Les chercheurs pouvaient ainsi mesurer la qualité de prévision obtenue sur des cas auxquels les algorithmes n'avaient pas été exposés. Ils utilisaient pour cela l'erreur quadratique moyenne (MSE), l'erreur de prédiction pour chaque cas étant le carré de l'écart entre la prédiction de l'algorithme et le résultat réel.

Quelle est donc cette qualité de prévision, pour les algorithmes gagnants ? Bien entendu, les algorithmes d'apprentissage automatique formés à partir d'un grand ensemble de données font mieux que des modèles linéaires simples (et, par extension, que des humains). Mais l'amélioration qu'apporte l'intelligence artificielle par rapport à un modèle très simple est somme toute modeste, et la justesse de ses prédictions reste, dans l'absolu, assez décevante. Quand il s'agit de prévoir si une famille sera expulsée de son domicile, par exemple, le meilleur modèle arrive à une corrélation de 0,22 (PC = 57 %)[3]. La qualité de prévision est très voisine quand il s'agit de prévoir si l'un des parents va être licencié, s'il va suivre une formation, ou si l'on cherche à estimer le score qu'obtiendra l'enfant sur des mesures de personnalité comme la persévérance dans la durée (*grit*). Pour ces résultats, les corrélations entre la prédiction du meilleur modèle et la réalité s'échelonnent entre 0,17 et 0,24 (PC = 55 à 58 %).

Seuls deux des six résultats cibles s'avèrent beaucoup plus prévisibles. Pour la moyenne générale de l'enfant au lycée, la corrélation prédictive atteinte par le meilleur modèle est de 0,44 (PC = 65 %). Pour un indice global des difficultés matérielles de la famille pendant les douze mois précédents, elle est de 0,48 (PC = 66 %). La raison de cette meilleure performance est purement technique : ces résultats sont eux-mêmes des

agrégats. La moyenne générale au lycée agrège les notes dans différentes matières. De même, l'indice de difficulté matérielle est une synthèse des réponses à onze questions différentes, parmi lesquelles : « Vous est-il arrivé d'avoir faim ? » ou « Votre ligne téléphonique a-t-elle été coupée ? ». Or les statisticiens savent que les agrégats sont plus prévisibles (et plus prédictifs) que les mesures ponctuelles.

En somme, la principale conclusion de ce concours de prévision est que même avec une masse imposante d'informations supposées prédictives, il reste extraordinairement difficile de prévoir des événements précis dans la vie des gens – et que même les prévisions portant sur des agrégats sont d'une exactitude assez limitée.

Ces résultats ne sont en vérité pas très surprenants : beaucoup de corrélations observées en sciences sociales sont du même ordre que celles rapportées ici. En 2003, une grande synthèse de travaux de recherche en psychologie sociale, couvrant vingt-cinq mille études qui avaient impliqué huit millions de sujets sur une centaine d'années, concluait que, « en psychologie sociale, la taille d'effet typique est un *r* [coefficient de corrélation] de 0,21[4] ». Des corrélations beaucoup plus élevées, comme le *r* de 0,60 que nous avons évoqué entre la taille et la pointure d'un homme adulte, sont fréquentes en matière de mesures physiques, mais très rares en sciences sociales. Une autre synthèse, portant sur sept cent huit études en sciences comportementales et cognitives, suggère que 3 % seulement des corrélations rapportées sont de 0,50 ou plus[5].

Ces coefficients de corrélation très modestes vous surprendront peut-être, tant il est fréquent d'entendre parler de résultats « statistiquement significatifs » ou même « hautement significatifs ». Les termes statistiques sont souvent trompeurs pour le profane, et le mot « significatif » en est peut-être le pire exemple. Quand un résultat est qualifié de « significatif », il ne faut pas en conclure que l'effet qu'il décrit est puissant. Cela veut simplement dire qu'il est peu probable que le résultat observé soit le produit

du hasard. Pour peu que l'échantillon soit suffisamment grand, une corrélation peut être à la fois « significative » au sens statistique et absolument insignifiante par la taille de ses effets.

L'étude de Princeton, avec ses conclusions troublantes sur l'imprévisibilité de la vie, invite à une réflexion sur la différence entre compréhension et prédiction. La base de données *Fragile Families* est considérée comme un véritable trésor des sciences sociales, et comme nous l'avons vu, elle a servi de base à un très abondant corpus de recherches. Les chercheurs qui ont produit ces articles ont certainement eu le sentiment que leurs travaux contribuaient à mieux *comprendre* la vie des familles fragiles. Malheureusement, ce sentiment de progrès ne s'accompagne pas d'une capacité à faire des *prévisions* granulaires sur des événements de vie individuels, même lorsque les données sont exploitées par les techniques les plus avancées pour produire la meilleure performance prédictive possible. Dans l'introduction de leur article, McLanahan, Salganik et leurs nombreux coauteurs résument le problème d'une phrase : « Les chercheurs vont devoir confronter leur sentiment de comprendre les trajectoires de vie avec le fait qu'aucune des prévisions qu'on peut en faire ne s'avère très juste[6]. »

## *Compréhension et prédiction*

La logique qui sous-tend cette conclusion pessimiste nécessite un peu d'explications. Quand les auteurs du concours de Princeton mettent sur le même plan la compréhension et la prédiction (ou l'absence de l'une et l'absence de l'autre), ils utilisent le mot *compréhension* dans un sens bien précis. Il y en a d'autres : quand vous dites que vous « comprenez » un concept mathématique, ou que vous « comprenez » ce qu'est l'amour, vous ne revendiquez sans doute pas une capacité à faire des prédictions précises.

Dans le discours des sciences sociales, cependant, et d'ailleurs dans la plupart des conversations quotidiennes, dire qu'on comprend une chose, c'est affirmer qu'on comprend ce qui *cause* cette chose. Les sociologues qui ont collecté et étudié les milliers de variables de l'étude *Fragile Families* étaient en quête de causes expliquant les résultats qu'ils observaient. Quand des médecins « comprennent » ce dont souffre un patient, ils veulent dire que la pathologie qu'ils ont diagnostiquée est la cause des symptômes observés. Comprendre, c'est décrire une chaîne causale[7]. La possibilité de faire une prédiction juste est une manière de vérifier que cette chaîne causale a bien été identifiée. Et la corrélation, qui mesure la justesse prédictive, est une manière de mesurer la causalité, de quantifier le lien entre la cause et l'effet qu'on tente d'expliquer.

Cette dernière assertion pourrait vous étonner si vous vous souvenez d'un avertissement souvent répété dans les cours d'introduction aux statistiques : « corrélation n'est pas causalité ». Prenons par exemple la corrélation entre la pointure de chaussures d'un enfant et ses capacités mathématiques : à l'évidence, aucune de ces deux variables n'est la cause de l'autre. La corrélation, bien réelle, vient du fait que la taille du pied et les connaissances mathématiques augmentent toutes deux avec l'âge de l'enfant. La corrélation permet une prédiction : si vous savez qu'un enfant a de grands pieds, vous pouvez lui prédire un niveau en maths supérieur à celui d'un enfant qui a des pieds plus petits. Mais vous ne devriez évidemment pas inférer pour autant un lien de causalité.

En règle générale, cependant, si la corrélation n'implique pas la causalité, la causalité, toutes choses égales par ailleurs, implique bien la corrélation. Là où il y a des liens de cause, il doit y avoir une corrélation. S'il n'y a pas de corrélation entre l'âge et la pointure chez les adultes, on peut conclure avec confiance qu'à partir de la fin de l'adolescence, les pieds ne grandissent plus, et qu'il faut chercher ailleurs les causes des différences de pointure.

Pour résumer : dans le contexte de cette étude, causalité implique corrélation. Il s'ensuit que là où il y a causalité, nous devrions être capables de faire une prédiction ; et que la corrélation, qui mesure la justesse de cette prédiction, mesure aussi ce que nous comprenons de la causalité. D'où la conclusion des chercheurs de Princeton : la capacité à prédire des événements de vie comme les expulsions, capacité chiffrée par une corrélation de 0,22, donne la mesure du niveau de compréhension – ou d'incompréhension – qu'ont les sociologues des parcours de vie. L'ignorance objective impose un plafond non seulement à nos prédictions, mais aussi à notre compréhension.

Dans ces conditions, que veulent dire la plupart des professionnels quand ils affirment avec confiance qu'ils « comprennent » leur domaine ? Comment peuvent-ils se prononcer sur ce qui cause les phénomènes qu'ils observent et faire des prédictions confiantes à leur sujet ? En bref, pourquoi ces professionnels, et nous tous avec eux, sous-estiment-ils tellement leur ignorance objective du monde ?

## *La pensée causale*

Peut-être, en lisant la première partie de ce chapitre, vous êtes-vous interrogé sur les facteurs qui pourraient expliquer les expulsions et les autres événements de vie chez les familles fragiles. Si c'est le cas, vous avez utilisé le même mode de pensée que les chercheurs dont nous décrivions les travaux : une *pensée statistique*. Vous vous êtes intéressé à des ensembles, par exemple à la population des familles fragiles, considérée comme un tout. Vous y avez réfléchi en termes statistiques, en pensant à des moyennes, à des variances, à des corrélations, etc. Vous n'avez pas songé à des cas individuels.

Mais il est fort possible que vous ayez utilisé un autre mode de pensée, qui vient plus naturellement à l'esprit : la *pensée causale*[8]. La pensée causale fabrique des histoires dans lesquelles des événements, des individus et des objets s'affectent les uns les autres.

Pour comprendre la pensée causale, imaginez par exemple que vous soyez un travailleur social qui suit un grand nombre de familles défavorisées. Vous venez d'apprendre qu'une de ces familles, les Jones, a été expulsée du domicile familial. Votre réaction à cet événement est nourrie par ce que vous savez des Jones. Il se trouve que Jessica Jones, qui subvient aux besoins de la famille, a été licenciée il y a quelques mois. Elle n'a pas pu retrouver d'emploi, et n'a pas pu depuis payer son loyer en totalité. Elle a fait des paiements partiels, plaidant sa cause à plusieurs reprises auprès du gérant de son immeuble, et vous a même demandé d'intervenir (ce que vous avez fait, mais sans émouvoir ce dernier). Dans ce contexte, l'expulsion des Jones est triste, mais pas étonnante. Elle semble être la conclusion logique d'une chaîne d'événements, le dénouement inévitable d'une tragédie déjà écrite.

Quand nous cédons à ce sentiment d'inévitabilité, nous oublions que les événements auraient pu être différents, et qu'à chaque tournant, le destin aurait pu emprunter une autre route. Jessica Jones aurait pu conserver son emploi. Elle aurait pu en retrouver un autre rapidement. Une parente aurait pu lui venir en aide. Vous, le travailleur social, auriez pu plaider plus efficacement sa cause. Le gérant de l'immeuble aurait pu se montrer plus compréhensif et donner à la famille quelques semaines de répit, ce qui aurait permis à Jessica de retrouver un emploi et de régler ses arriérés de loyer.

Ces récits alternatifs sont aussi peu surprenants que le premier – à condition qu'on en connaisse la fin. Quel que soit le résultat (l'expulsion ou non), une fois qu'il est arrivé, la pensée causale donne l'impression non seulement qu'il est totalement explicable, mais même qu'il était prévisible.

## *La compréhension dans la vallée des Choses Normales*

Il y a une explication psychologique à cette observation. Certains événements sont surprenants : une pandémie terrible, un attentat contre les tours jumelles du World Trade Center, la chute d'un grand fonds d'investissement dont le gérant était en fait un escroc. De la même manière, notre vie personnelle peut aussi nous réserver de grandes surprises, comme un coup de foudre amoureux, la mort soudaine d'un enfant, ou un héritage inattendu.

D'autres événements, quant à eux, sont activement attendus. On peut ainsi attendre activement le résultat d'un match de football ou l'arrivée d'un écolier qui doit rentrer à la maison à une heure précise.

Mais l'expérience humaine se situe en général entre ces deux extrêmes. Bien sûr, nous sommes parfois surpris, et nous sommes quelquefois dans l'attente vigilante d'un événement précis. Mais la plupart des choses se produisent dans la vaste vallée qui sépare ces deux extrêmes : la vallée des Choses Normales. Dans la vallée des Choses Normales, les événements surviennent naturellement. Ils ne sont ni explicitement attendus, ni particulièrement surprenants. À cet instant, par exemple, vous n'avez pas d'attentes précises sur ce que contiendra le prochain paragraphe de ce livre. Vous seriez très étonné qu'il soit écrit en turc, bien sûr, mais il y a énormément de choses que nous pourrions y exprimer sans que vous en soyez surpris.

Dans la vallée des Choses Normales, les événements s'enchaînent comme ceux qui ont conduit à l'expulsion des Jones : ils semblent « normaux » après coup, alors même qu'ils n'étaient pas activement attendus et que nous n'aurions pas été capables de les prédire. Il en va ainsi parce que le processus par lequel nous comprenons la réalité consiste à regarder dans le rétroviseur[9]. Un événement qui n'a pas été activement

anticipé, comme l'expulsion de la famille Jones, déclenche un travail de mémoire qui va rechercher dans le passé une cause possible : les difficultés du marché de l'emploi, le gérant intransigeant. La recherche s'arrête sitôt qu'un récit plausible a été trouvé. Si le résultat avait été opposé, la recherche aurait produit des causes tout aussi convaincantes : la formidable ténacité de Mme Jones, le gérant compréhensif.

Comme le montrent ces exemples, beaucoup d'événements dans une histoire ordinaire s'expliquent littéralement tout seuls. Peut-être avez-vous noté que le gérant de l'immeuble n'était pas la même personne dans les deux versions de l'histoire. Mais le seul indice que vous avez sur la personne du gérant est son comportement. Compte tenu de ce que nous savons de lui, son comportement paraît cohérent, parce que ce que nous savons de lui provient justement du résultat que nous cherchons à expliquer. C'est l'occurrence de l'événement qui nous en indique la cause.

Quand un résultat est inattendu mais pas surprenant, et quand on l'explique de cette façon, on parvient toujours à lui trouver un sens. C'est cela que nous voulons dire quand nous disons que nous *comprenons* une histoire. Et c'est ce qui fait que la réalité paraît prévisible – après coup, quand nous la regardons dans le rétroviseur. Comme l'événement s'explique lui-même en se produisant, nous entretenons aisément l'illusion que nous aurions pu l'anticiper.

Plus largement, le sentiment que nous avons de comprendre le monde repose sur notre extraordinaire aptitude à bâtir des récits qui expliquent ce que nous observons. La recherche des causes arrive presque toujours à ses fins, parce que le réservoir de faits et de croyances dans lequel nous pouvons aller les piocher est pratiquement illimité. Il suffit par exemple d'écouter les informations du soir pour vérifier qu'aucun mouvement boursier important ne reste jamais inexpliqué. On invoque volontiers l'actualité pour expliquer une baisse des indices (« après les événements récents, les gérants sont inquiets »). Mais la même actualité peut tout aussi

facilement « expliquer » la hausse des mêmes indices (« convaincus que les banques centrales sauront réagir aux événements récents, les gérants gardent leur sang-froid »).

Que faisons-nous quand la recherche d'une cause évidente n'aboutit pas ? Notre premier recours, alors, est de produire une explication en complétant notre modèle du monde. Ce faisant, nous remplissons un « blanc » en inférant un fait dont nous n'avions jusque-là pas connaissance et sur lequel nous ne nous étions pas interrogés : nous concluons, par exemple, que le gérant de l'immeuble est d'une exceptionnelle humanité. C'est seulement quand notre modèle du monde ne peut pas être bricolé pour produire le résultat observé que nous qualifions celui-ci de surprenant ; et c'est seulement dans ce cas que nous nous mettons à chercher une explication plus élaborée. Il n'y a d'étonnement véritable que quand notre démarche habituelle d'explication rétrospective échoue.

C'est cette lecture causale et continuellement réinterprétée de la réalité qui constitue notre « compréhension » du monde. Dans la vallée des Choses Normales, où le flot de la vie s'écoule paisiblement, il nous suffit de nous tourner vers le passé pour éprouver le sentiment puissant de comprendre les événements qu'il nous a apportés. Ce sentiment est fondamentalement causal : une fois connus, les événements nouveaux éliminent les autres possibilités, et le récit laisse peu de place au doute. Nous pouvons bien sûr éprouver brièvement un sentiment d'incertitude ; mais, comme nous l'enseignent les travaux de recherche classiques sur la vision rétrospective, le souvenir de ce sentiment s'évanouit dès que l'incertitude s'est dissipée [10].

## *Pensée statistique, vision externe*

Nous avons mis en évidence deux manières radicalement différentes de penser les événements : la pensée statistique et la pensée causale.

Le mode causal nous permet une grande économie de moyens : en catégorisant les événements en temps réel comme normaux ou anormaux, nous nous épargnons l'effort de réfléchir à tout ce qui ne nous semble pas remarquable. Seuls les événements étiquetés comme anormaux mobilisent notre attention et déclenchent une recherche d'informations, dans la réalité ou dans notre mémoire. L'attente active – quand nous attendons avec vigilance que quelque chose se produise – exige elle aussi un effort. En revanche, l'écoulement des événements dans la vallée des Choses Normales nous sollicite très peu. Le voisin que vous croisez peut vous sourire ou se contenter d'un hochement de tête : ni l'un ni l'autre de ces événements, s'il s'est produit assez souvent par le passé, n'attirera votre attention. C'est seulement si le sourire est inhabituellement chaleureux ou le hochement de tête inhabituellement brusque que vous commencerez à lui chercher dans vos souvenirs une cause possible. La pensée causale évite tout effort superflu, tout en conservant la vigilance nécessaire pour détecter un événement anormal.

La pensée statistique demande, au contraire, beaucoup d'efforts. Elle requiert une attention que seul le Système 2, le mode de pensée associé à la réflexion lente et délibérée, peut mobiliser[11]. De surcroît, au-delà d'un niveau élémentaire, la réflexion statistique exige des connaissances spécifiques. La pensée statistique s'attache à des ensembles : les cas individuels ne sont pour elle que des exemples de catégories plus larges. Ainsi, si l'on pense statistique, l'expulsion de la famille Jones n'est pas le résultat d'une chaîne d'événements spécifiques : c'est un résultat statistiquement probable (ou improbable), compte tenu des observations antérieures d'un ensemble de cas présentant, sur des caractéristiques considérées comme prédictives, des similitudes avec celui des Jones.

Le contraste entre ces deux perspectives – causale et statistique – est un des thèmes récurrents de ce livre. Appliquer la pensée causale à un cas particulier est une source d'erreurs prévisibles. Recourir à la perspective statistique, que nous appellerons aussi la *vision externe*, est un moyen d'éviter ces erreurs.

À ce stade, le point central qu'il faut souligner est que le mode de pensée causal nous est beaucoup plus naturel. Même des explications qu'on devrait naturellement considérer comme statistiques sont facilement transformées en récits causaux. Songez à des réflexions comme : « Ils ont échoué parce qu'ils manquaient d'expérience » ou « Ils ont réussi parce qu'ils avaient un chef brillant ». Il ne vous serait pas difficile de trouver des contre-exemples, dans lesquels des équipes peu expérimentées ont réussi et des chefs brillants ont échoué. Statistiquement parlant, les corrélations de l'expérience et de la brillance avec la réussite sont au mieux modestes et probablement faibles. Pourtant, une explication causale est vite trouvée. Il suffit qu'une causalité nous semble plausible pour que notre esprit traite une corrélation faible comme une propension dotée de puissance explicative. Un leadership brillant est accepté comme une explication suffisante du succès, l'inexpérience comme une explication satisfaisante de l'échec.

Sans doute ce mode d'explication, pour erroné qu'il soit, est-il inévitable : personne ne voudrait renoncer au sentiment de comprendre le monde. Le problème est que la pensée causale, en nous donnant l'illusion de comprendre le passé, contribue aussi à ce que nous produisions des prédictions excessivement confiantes de l'avenir. En outre, comme nous le verrons, la préférence pour la pensée causale nous masque la perception du bruit comme source d'erreur, parce que le bruit est une notion fondamentalement statistique.

---

Le monde est loin d'être aussi prévisible que nous le pensons. Dans beaucoup de domaines, et notamment en sciences sociales, même les outils

les plus puissants – sans parler de notre jugement – peinent à formuler des prévisions pertinentes. Si nous éprouvons néanmoins le sentiment de « comprendre » les événements, c'est parce que la pensée causale nous permet de leur trouver un sens.

Ainsi, nous jugeons les événements beaucoup plus prévisibles qu'ils ne le sont en réalité. Dans la vallée des Choses Normales, il n'y a ni surprises, ni incohérences. L'avenir semble aussi prévisible que le passé. Et le bruit reste inaudible.

### À propos des limites de la compréhension

« Les corrélations d'environ 0,20 (PC = 56 %) sont assez fréquentes dans les affaires humaines. »

« La corrélation n'implique pas la causalité, mais la causalité implique la corrélation. »

« Dans la vallée des Choses Normales, les événements ne sont ni attendus ni surprenants : leur survenance suffit à les expliquer. »

« Nous croyons comprendre ce qui se passe. Mais aurions-nous pu le prévoir ? »

PARTIE IV

# D’où vient le bruit ?

D’où vient le bruit – et d’où vient le biais ? Quels sont les mécanismes mentaux qui introduisent de la variabilité dans nos jugements et qui provoquent des erreurs communément partagées ? Telles sont les questions que nous allons maintenant examiner.

Pour explorer la psychologie du bruit, nous commencerons, dans le chapitre 13, par décrire certaines opérations de la pensée rapide du « Système 1 » qui sont responsables de nombreuses erreurs de jugement.

Dans les trois chapitres suivants, nous nous pencherons sur trois mécanismes générateurs de bruit : la création d’équivalences entre des jugements sur des dimensions différentes, le choix des échelles sur lesquelles les jugements sont produits, et le rôle des personnalités, génératrices de patterns de jugements uniques et donc différents.

Dans le chapitre 17, enfin, nous passerons en revue ce que nous aurons appris jusque-là sur le bruit et sur ses composantes. Cela nous amènera à compléter notre réponse à une question que nous avons déjà soulevée : pourquoi le bruit, malgré son omniprésence, est-il rarement considéré comme un problème sérieux ?

CHAPITRE 13

# Heuristiques, biais et bruit

Ce livre prolonge un demi-siècle de travaux théoriques et expérimentaux sur le jugement, formant un courant de recherche qu'il est convenu d'appeler les « heuristiques et biais ». Les quatre premières décennies de ces travaux sont résumées dans *Système 1/Système 2*, qui explore les mécanismes psychologiques expliquant les merveilleuses aptitudes de la pensée intuitive, mais aussi ses failles[1].

L'idée directrice de la recherche sur les heuristiques et biais est que les gens à qui l'on pose une question ardue recourent à des opérations simplificatrices appelées *heuristiques*. En général, ces heuristiques, qui sont produites par la pensée intuitive et rapide, celle du Système 1, sont fort utiles et produisent des réponses satisfaisantes. Mais il arrive quelquefois qu'elles conduisent à des biais, c'est-à-dire à des erreurs de jugement dont l'occurrence et la direction sont prévisibles.

La recherche sur les heuristiques et les biais ne s'intéressait pas prioritairement à ce qui rend les gens différents les uns des autres, mais bien à ce qu'ils ont en commun. Sa conclusion la plus connue est que certains processus mentaux largement partagés causent des erreurs de jugement qui le sont également. En partie pour cette raison, les lecteurs familiers de la notion de *biais* au sens *psychologique* du terme croient souvent que les

erreurs qui en résultent sont toujours un *biais statistique*, au sens que nous utilisons dans ce livre : un jugement qui s'écarte de la vérité, et qui s'en écarte de la même manière et dans la même direction chez la plupart des gens.

Il y a là un malentendu qu'il faut dissiper. Pour ce faire, nous allons nous attacher dans ce chapitre à comprendre les effets des biais psychologiques. Indéniablement, ceux-ci peuvent être sources de biais statistique : c'est le cas lorsqu'ils sont largement partagés. Cependant, dès lors que plusieurs juges sont sujets à des biais différents (ou sujets aux mêmes biais mais dans des proportions différentes), leurs biais psychologiques ne créent pas une erreur identique : ils créent de la variabilité entre les juges, donc du bruit systémique. Dans les deux cas, bien sûr, les biais psychologiques créent de l'erreur ; mais pas de même nature.

## *Diagnostiquer les biais*

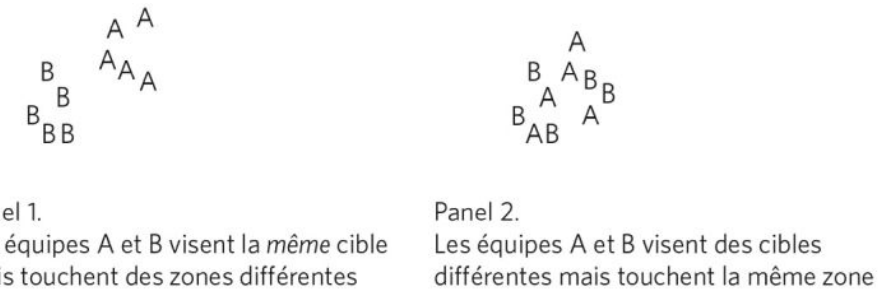
On identifie souvent les biais dans le jugement par comparaison avec une valeur juste ou vraie. Il y a par exemple un biais dans un jugement prédictif quand la plupart des erreurs de prévision vont dans le même sens. Ainsi, quand des personnes évaluent le temps qu'il faudra pour mener à bien un projet, la moyenne de leurs estimations est généralement très en dessous du temps que le projet demandera réellement. Ce biais psychologique bien connu est appelé *erreur de prévision* ou *biais du planificateur*.

Il est cependant fréquent qu'il n'y ait pas de valeur vraie à laquelle les jugements puissent être comparés, puisque certains jugements, nous l'avons vu, sont invérifiables. Or nous avons défini un biais statistique comme un écart entre la moyenne des jugements et la valeur vraie. Nous avons par conséquent souligné qu'un tel biais ne peut être détecté que quand on connaît cette valeur vraie. Peut-être vous demandez-vous donc comment on

peut étudier les biais psychologiques sur des questions dont la réponse vraie n’est pas connue.

La réponse fait appel à une méthode indirecte : les chercheurs concluent à la présence d’un biais psychologique quand ils observent qu’un facteur qui *ne devrait pas* affecter le jugement a un effet statistique sur celui-ci, ou au contraire qu’un facteur qui *devrait* affecter le jugement n’a pas d’effet perceptible.

Pour illustrer cette méthode, revenons un instant au stand de tir de notre introduction. Imaginons que les équipes A et B aient fini de tirer et que nous examinions le dos de la cible (figure 12). Dans cette situation, vous ne savez pas où se trouve le centre de la cible : sa position, qui joue le rôle de la « valeur vraie », est inconnue. Vous ignorez donc dans quelle mesure le centre de gravité des tirs de chaque équipe s’écarte du centre de la cible, écart qui constitue l’erreur moyenne – le biais – de chaque équipe. Supposez, cependant, qu’on vous donne une précision importante : dans le panel 1, le centre de la cible visée était le même pour les deux équipes ; tandis que dans le panel 2, les deux équipes ne visaient pas le même point.




Panel 1.
Les équipes A et B visent la *même* cible mais touchent des zones différentes

Panel 2.
Les équipes A et B visent des cibles différentes mais touchent la même zone

**Figure 12.** Diagnostiquer un biais en regardant « le dos de la cible ».

Il est désormais possible, même sans connaître la position exacte du centre de la cible, de conclure à la présence de biais. Dans le panel 1, les tirs des deux équipes diffèrent, alors qu’ils devraient être identiques. Ce schéma

ressemble à ce que nous observerions, par exemple, si nous soumettions à deux groupes d'investisseurs en capital-risque des plans d'affaires identiques par leur contenu, mais imprimés dans des polices de caractères différentes sur des types de papier différents. Dans la mesure où ces détails, qui ne sont pas pertinents, modifient le jugement des investisseurs, nous pouvons conclure à la présence d'un biais psychologique. Nous ne savons pas si les investisseurs impressionnés par l'élégance de la police et la brillance du papier sont trop optimistes par rapport à la réalité des perspectives de la jeune pousse, ou si ceux qui ont lu la version plus médiocrement présentée sont trop pessimistes. Mais nous savons que leurs jugements sont différents, alors qu'ils ne devraient pas l'être.

Le panel 2 illustre le problème inverse : une absence d'écarts là où l'on devrait en trouver. Imaginons par exemple que l'on interroge deux groupes de personnes sur Michael Gambardi, le nouveau P-DG dont vous avez estimé les chances de succès au chapitre 4 – mais en faisant varier un détail. Au premier groupe, on demande – comme à vous – d'estimer la probabilité que Gambardi soit encore en poste dans deux ans. Au second, on demande la probabilité qu'il le soit encore dans trois ans. En bonne logique, les deux groupes devraient arriver à des conclusions différentes, car il y a évidemment plus de raisons et de manières de perdre son emploi en trois ans qu'en deux. Pourtant, sur la base de nombreuses études portant sur des problèmes analogues, il y a fort à parier que les estimations de probabilité des deux groupes différeront très peu, voire pas du tout[2]. Les réponses devraient être nettement différentes, mais elles ne le sont pas, ce qui suggère qu'un facteur quantitatif est ignoré alors qu'il devrait influencer les jugements. (Le biais psychologique en cause ici est appelé l'*insensibilité à l'étendue*.)

L'existence de ce type d'erreurs prévisibles et directionnelles a été démontrée sur de très nombreux problèmes, et le mot *biais* est désormais entré dans le langage courant de l'analyse économique, politique, juridique,

et de bien d'autres domaines. Le sens qu'on lui donne couramment est d'ailleurs bien plus large que la définition strictement cognitive que nous utilisons ici, et qui fait référence à un mécanisme psychologique et à l'erreur que ce mécanisme produit. Ainsi, on lira souvent qu'un individu est « biaisé » à l'encontre de certains groupes de personnes (on parle par exemple de biais de genre ou de biais ethnique). Le mot « biais » peut aussi signifier qu'un individu privilégie *a priori* une conclusion particulière : il est biaisé par un conflit d'intérêts ou par son adhésion à certaines thèses politiques. Nous incluons ces types de biais dans l'analyse que nous faisons ici de la psychologie des erreurs de jugement, parce que tous ces biais psychologiques peuvent produire aussi bien du biais statistique que du bruit.

Il existe toutefois un mauvais usage du mot « biais » : celui qui consiste à imputer un échec à la présence de « biais », sans plus de précision. Tout en reconnaissant l'erreur, on promet alors souvent de « tout faire pour éliminer les biais dans notre processus de décision ». Ce genre de propos n'ajoute rien aux constats convenus selon lesquels « des erreurs ont été commises » et « nous allons en tirer toutes les leçons pour nous efforcer de mieux faire à l'avenir ». Assurément, certains échecs sont réellement causés par des erreurs prévisibles, elles-mêmes associées à certains biais psychologiques, et il est bien sûr possible d'intervenir concrètement pour réduire le biais (et le bruit) dans les jugements et les décisions. Mais il est tout à fait stérile de mettre chaque résultat décevant sur le compte de vagues « biais ». Pour que le mot *biais* conserve un sens, il faut le réserver à des erreurs identifiables et bien précises et aux mécanismes qui les produisent.

## *Les biais de substitution*

Pour illustrer ce qu'est une heuristique, nous vous proposons une expérience qui met en scène plusieurs thèmes essentiels de la théorie des

heuristiques et biais. Ici encore, vous tirerez un plus grand profit de l'exemple si vous prenez quelques instants pour réfléchir à vos propres réponses.

*Bill a 33 ans. Il est intelligent mais sans imagination, compulsif et, de façon générale, assez amorphe. À l'école, il était fort en mathématiques mais faible dans les autres matières, en particulier littéraires.*

*Voici une liste de huit possibilités pour décrire la situation actuelle de Bill.*

*Choisissez les deux possibilités que vous jugez les plus probables.*

❑ *Bill est médecin et joue au poker pour ses loisirs.*
❑ *Bill est architecte.*
❑ *Bill est comptable.*
❑ *Bill joue du jazz pour ses loisirs.*
❑ *Bill fait du surf pour ses loisirs.*
❑ *Bill est reporter.*
❑ *Bill est comptable et joue du jazz pour ses loisirs.*
❑ *Bill fait de l'alpinisme pour ses loisirs.*

*Revenez maintenant à la liste et choisissez les deux catégories pour lesquelles Bill ressemble le plus à une personne représentative de ces catégories. Votre choix peut être identique ou différent du précédent.*

Vous avez presque certainement choisi les mêmes catégories pour les deux questions, c'est-à-dire pour évaluer la *probabilité* et la *ressemblance*. Si nous en sommes si sûrs, c'est parce que de nombreuses expériences ont démontré que les gens répondent de manière identique aux deux questions[3].

Pourtant, la ressemblance et la probabilité sont deux choses bien différentes. Demandez-vous, par exemple, laquelle des deux affirmations suivantes a le plus de sens :

❑ *Bill correspond à l'idée que je me fais d'une personne qui joue du jazz pour ses loisirs.*
❑ *Bill correspond à l'idée que je me fais d'un comptable qui joue du jazz pour ses loisirs.*

Aucune de ces affirmations n'est entièrement satisfaisante, mais il y en a une qui est moins aberrante que l'autre. Bill a plus de points communs avec un *comptable* qui fait du jazz pour ses loisirs qu'avec un individu lambda qui fait du jazz pour ses loisirs.

Mais demandez-vous maintenant laquelle des deux affirmations suivantes est la plus *probable* :

❑ *Bill joue du jazz pour ses loisirs.*
❑ *Bill est un comptable qui joue du jazz pour ses loisirs.*

Sans doute êtes-vous tenté, ici encore, de choisir la seconde réponse. Hélas, la logique l'interdit. La probabilité que Bill joue du jazz est nécessairement plus élevée que celle qu'il soit comptable *et* joue du jazz. Rappelez-vous vos diagrammes de Venn, ces « patates » qui aident à représenter le recouvrement de deux ensembles : l'ensemble des comptables qui jouent du jazz est entièrement inclus dans l'ensemble des personnes qui jouent du jazz. Par conséquent, si Bill est joueur de jazz et comptable, alors il est nécessairement joueur de jazz : la probabilité qu'il remplisse deux conditions ne peut pas être supérieure à la probabilité qu'il en remplisse une seule. L'ajout d'un détail à une description ne peut que rendre cette description moins probable, même s'il la rend en même temps plus représentative et donc plus « ressemblante », comme dans le cas présent.

Cet exemple illustre une thèse centrale de la théorie des heuristiques de jugement : les gens confrontés à une question difficile vont souvent lui substituer une question plus facile. Des deux questions suivantes, quelle est

la plus facile : « Dans quelle mesure Bill *ressemble-t-il* à quelqu'un qui joue du jazz ? » ou « Quelle est la *probabilité* que Bill joue du jazz en amateur ? ». De l'avis général, la question de la ressemblance est la plus facile. C'est pourquoi, en réalité, c'est sans doute à celle-ci que vous avez répondu, même quand la question portait en fait sur la probabilité.

Malheureusement, la substitution d'une question à une autre provoque des erreurs prévisibles, qui sont des exemples de biais psychologiques. L'exemple de Bill en fournit une excellente démonstration. Lorsqu'on substitue un jugement de ressemblance à un jugement de probabilité, il est inévitable que des erreurs surviennent, car les jugements de probabilité sont soumis à des règles logiques qui leur sont propres. Les diagrammes de Venn, en particulier, s'appliquent aux probabilités, pas aux ressemblances. D'où l'erreur logique prévisible que l'on fait souvent.

Il existe de nombreux autres exemples de propriétés statistiques qu'on a tendance à négliger. Le problème Gambardi, que vous avez rencontré au chapitre 4, en illustre une. Rappelez-vous la manière dont vous avez raisonné : selon toute vraisemblance, l'évaluation que vous avez faite des chances de réussite de Michael Gambardi reposait entièrement sur les informations que le texte vous donnait à son sujet. Vous vous êtes ensuite demandé dans quelle mesure cette description correspondait à votre image mentale d'un P-DG qui réussit.

Mais il y a une question que vous ne vous êtes très probablement pas posée : « Quelle est la probabilité qu'un nouveau P-DG *sélectionné au hasard* soit toujours en poste deux ans plus tard ? » Nous pouvons considérer ce *taux de base* comme une mesure de la difficulté qu'il y a, *en général*, à conserver deux ans durant un poste de P-DG.

Si ce raisonnement vous semble étrange, songez à la manière dont vous estimeriez la probabilité qu'une étudiante donnée réussisse un examen. Bien sûr, il est utile de connaître les qualités de l'étudiante en question ; mais le taux de succès à l'examen est un élément capital, parce qu'il vous

donne une indication de sa difficulté. De la même façon, il est fort utile, pour estimer les chances de Michael Gambardi, de connaître le taux de base de « survie à deux ans » d'un P-DG. Cette logique illustre ce que nous avons appelé la vision externe : quand on adopte une vision externe, on considère Gambardi (ou l'étudiante) comme des éléments dans un ensemble de cas similaires. Au lieu d'appliquer la pensée causale à un cas particulier, on utilise la pensée statistique pour considérer une catégorie dans son ensemble.

La vision externe peut prévenir des erreurs considérables. Quelques minutes de recherches suffisent pour trouver une estimation du taux annuel de turnover des P-DG dans les entreprises américaines : environ 15 %[4]. Un P-DG moyen qui est en poste aujourd'hui a donc environ 72 % de chances de l'être encore dans deux ans. Bien sûr, ce chiffre n'est qu'un point de départ, et les nombreuses particularités du cas Gambardi vont affecter votre estimation finale. Mais si vous vous êtes contenté de ce que l'on vous a dit de Gambardi, vous avez négligé une information essentielle – comme presque tout le monde. (Permettez-nous un aveu à ce propos : nous avons conçu le cas Gambardi pour illustrer le bruit dans le jugement, et il nous a fallu plusieurs semaines pour nous apercevoir que c'était aussi un parfait exemple du biais que nous décrivons ici, appelé la *négligence du taux de base*. Penser aux taux de base ne vient pas spontanément à l'esprit, et pas plus pour les auteurs de ce livre que pour n'importe qui.)

L'utilisation de la ressemblance en lieu et place de la probabilité n'est pas le seul exemple de substitution d'une question à une autre. Une autre substitution habituelle consiste à remplacer un jugement de *fréquence* par une évaluation de la facilité avec laquelle un cas similaire vient à l'esprit. Ainsi, la perception du risque d'un accident d'avion ou d'un ouragan augmente brièvement quand ce type d'événement vient de se produire et qu'il lui a été donné une large publicité. En réalité, des événements qui sont arrivés récemment ne sont pas plus fréquents : ils sont simplement perçus

comme tels parce qu'ils sont plus « disponibles » : ils viennent plus facilement à l'esprit. L'heuristique qui procède à cette substitution est appelée *heuristique de disponibilité*.

Il y a bien d'autres types de jugements où l'on substitue un jugement facile à un jugement difficile. Répondre à une question plus simple est une sorte de moyen universel de répondre à une question à laquelle on ne trouve pas de réponse. Considérez par exemple chacune des questions suivantes, et la question de substitution bien plus facile que nous avons tendance à utiliser à sa place :

*Est-ce que je crois au changement climatique ?*

Est-ce que je fais confiance aux gens qui disent que le climat change ?

*Est-ce que je pense que ce chirurgien est compétent ?*

Ce chirurgien s'exprime-t-il avec assurance et autorité ?

*Le projet sera-t-il réalisé dans les délais ?*

Le projet est-il dans les temps aujourd'hui ?

*L'énergie nucléaire est-elle nécessaire ?*

Le mot « nucléaire » m'inspire-t-il de l'aversion ?

*Suis-je satisfait de ma vie en général ?*

Suis-je de bonne humeur à ce moment précis ?

Quelle que soit la question posée, lui en substituer une autre conduira inévitablement à une pondération erronée des données du problème, et cette pondération incorrecte produira de l'erreur. Il est clair, par exemple, qu'il ne suffit pas de consulter votre humeur du moment pour savoir si votre vie est satisfaisante ; mais les études sur cette question démontrent que la plupart des gens qui y répondent surpondèrent leur humeur passagère. On fait aussi

une erreur de pondération quand on substitue la ressemblance à la probabilité : on attribue alors un poids nul aux taux de base, qui sont sans intérêt pour juger de la ressemblance, mais qui seraient essentiels pour évaluer la probabilité.

De même, mais en sens inverse, quand des facteurs non pertinents (comme la qualité esthétique d'un plan d'affaires) affectent l'estimation de la valeur d'une entreprise, ces facteurs sont affectés d'un poids significatif alors qu'ils ne devraient pas influencer le jugement. Dans tous ces exemples, la substitution entraîne une pondération erronée des informations disponibles, et cette erreur de pondération produit une erreur de jugement.

## *Les biais de conclusion*

Lors d'un moment clé de l'écriture du scénario du *Retour du Jedi*[5], le troisième film de la saga *Star Wars*, George Lucas, le créateur de la série, eut un vif débat avec son bras droit, le scénariste Lawrence Kasdan, qui lui donna avec insistance ce conseil : « Je pense que tu devrais tuer Luke, et laisser Leia prendre les choses en main. » Lucas rejeta aussitôt cette idée. Kasdan suggéra alors que si Luke restait en vie, un autre personnage important devait mourir. Lucas n'était pas d'accord : « On ne tue pas les gens comme ça. » Kasdan répondit en invoquant la nature du cinéma lui-même. Il expliqua que le film « a davantage de poids émotionnel quand on perd quelqu'un qu'on aime ; l'odyssée a plus d'impact ».

La réponse de Lucas fut rapide et sans équivoque : « Je n'aime pas cette idée et je n'y crois pas. »

Le processus de pensée à l'œuvre ici est bien différent de celui que vous avez utilisé quand vous avez répondu aux questions sur Bill, le comptable amateur de jazz. Relisez la réponse de Lucas : *ne pas aimer* précède *ne pas y croire*. Lucas a réagi de manière automatique à la question de Kasdan.

Cette réaction est la principale raison de son jugement (même si celui-ci, aux yeux des fans de *Star Wars*, s'est avéré excellent).

Cet exemple de jugement hâtif, ou *préjugement*, illustre un autre type de biais, que nous appelons les *biais de conclusion*. Comme George Lucas, nous sommes souvent enclins, dès le début du processus de jugement, à favoriser une certaine conclusion. Quand c'est le cas, nous laissons notre Système 1, avec sa pensée intuitive et rapide, nous suggérer cette conclusion. Nous pouvons alors parfois nous précipiter sur cette conclusion en court-circuitant purement et simplement le processus de recueil et de traitement de l'information ; mais souvent, comme Lucas dans notre exemple, nous n'en restons pas là : nous mobilisons le Système 2 – la pensée réfléchie – pour aller chercher des arguments permettant d'étayer notre préjugement. Les arguments en question seront à la fois sélectifs et déformés[6] : nous aurons tendance à recueillir et à interpréter les informations de façon sélective pour justifier un jugement auquel nous croyons déjà (c'est le *biais de confirmation*) ou dont nous souhaitons qu'il soit vrai (c'est le *biais de désirabilité*).

Ce processus permet souvent de justifier nos préjugements en leur trouvant des rationalisations plausibles. Nous finissons même souvent par nous convaincre sincèrement que ces rationalisations sont la cause de nos croyances. Une bonne manière de déterminer si vous êtes victime d'un préjugement est d'imaginer que les arguments qui soutiennent votre conviction se révèlent soudain invalides. Kasdan, par exemple, aurait pu aisément démontrer à Lucas que l'argument selon lequel « on ne tue pas les gens comme ça » est un peu mince. *Roméo et Juliette* n'existerait pas si Shakespeare avait été de cet avis ; et si les scénaristes des *Sopranos* et de *Game of Thrones* s'étaient interdit de supprimer des personnages, leurs séries n'auraient sans doute pas dépassé la première saison. Mais on peut parier que ces objections n'auraient pas fait changer Lucas d'avis. Il aurait

simplement trouvé une autre raison pour justifier son jugement. Par exemple : « D'accord, mais *Star Wars*, c'est différent. »

Les préjugements sont omniprésents. Comme la réaction de Lucas, ils ont souvent une dimension émotionnelle. Le psychologue Paul Slovic parle à ce propos d'*heuristique de l'affect* : les gens décident de ce qu'ils croient en fonction de ce qu'ils ressentent. Chez les personnes que nous admirons profondément, nous aimons presque tout ; mais si nous pensons à un homme ou une femme politique dont nous combattons les idées, même son physique ou sa voix nous inspirent de l'aversion. C'est une des raisons pour lesquelles les entreprises se donnent tant de mal pour associer un affect positif à leur marque. Les professeurs remarquent souvent que les années où leur enseignement est bien noté, les étudiants donnent aussi de bonnes notes aux livres dont la lecture est recommandée pour le cours. Si l'année suivante, les étudiants n'apprécient pas le professeur, ils n'apprécient pas non plus les ouvrages conseillés – qui sont bien sûr restés strictement les mêmes. Même quand aucune émotion n'entre en jeu, un mécanisme analogue est à l'œuvre : quelles que soient les raisons véritables de votre conviction, vous serez enclin à accepter tout argument qui semble l'étayer, même si le raisonnement sous-jacent est erroné[7].

Il existe un exemple plus subtil de biais de conclusion : le *biais d'ancrage*, c'est-à-dire l'effet qu'un chiffre arbitraire a sur les gens qui formulent un jugement quantitatif. Une expérience classique consiste à vous présenter un certain nombre d'articles dont le prix est difficile à deviner : par exemple, une bouteille d'un vin que vous ne connaissez pas[8]. On vous demande ensuite de noter sur un papier les deux derniers chiffres de votre numéro de Sécurité sociale, qui composent un nombre compris entre 00 et 99, et d'indiquer si vous seriez prêt à payer ce prix, exprimé en dollars, pour la bouteille de vin. Enfin, on vous demande de dire quelle somme maximale vous seriez prêt à débourser pour cette même bouteille. L'analyse des résultats montre que votre numéro de Sécurité sociale influence le prix final

que vous indiquez. Ainsi, les gens « ancrés » par leur numéro de Sécurité sociale sur un prix élevé (plus de 80 dollars) se disaient prêts à payer en moyenne trois fois plus cher pour la bouteille de vin que ceux dont le numéro de Sécurité sociale était inférieur à 20.

Il paraît évident que votre numéro de Sécurité sociale ne devrait pas affecter votre estimation de la valeur d'une bouteille de vin, et pourtant il l'affecte – dans des proportions considérables. L'effet d'ancrage se retrouve avec une grande régularité, si bien qu'il est souvent utilisé à dessein, notamment dans les négociations[9]. Que vous marchandiez aux puces ou que vous participiez à une transaction complexe, vous aurez probablement intérêt, quand on en viendra aux chiffres, à parler le premier et à « ancrer » la discussion sur votre chiffre. La partie adverse sera en effet involontairement encline à chercher des justifications à votre offre, des raisons de la trouver raisonnable. Nous essayons toujours de trouver du sens à ce que nous entendons, si bien que, même quand un chiffre ne nous semble pas correct, nous cherchons automatiquement ce qui serait susceptible de réduire cette invraisemblance.

## *L'excès de cohérence*

Pour faire l'expérience d'un troisième type de biais, voici un autre petit exercice. Vous allez lire la description d'un candidat à un poste de cadre dirigeant. La description consiste en quatre adjectifs, chacun écrit sur une carte. Le jeu de cartes vient d'être battu. Les deux premières cartes que vous tirez indiquent deux adjectifs :

*Intelligent, Persévérant.*

Il serait raisonnable de suspendre votre jugement tant que vous ne disposez pas de toutes les informations, mais ce n'est pas ce qui s'est passé : vous avez déjà procédé à une évaluation du candidat, et elle est positive. Un jugement, simplement, s'est produit. Vous n'avez pas contrôlé le processus : vous retenir de conclure n'était pas une option.

Vous tirez ensuite les deux cartes suivantes. La description du candidat est maintenant la suivante :

*Intelligent, Persévérant, Rusé, Sans scrupule.*

Votre évaluation n'est plus favorable, mais elle n'a pas changé suffisamment. Pour vous en rendre compte, imaginez ce qu'aurait pu produire un autre ordre de tirage des mêmes cartes :

*Sans scrupule, Rusé, Persévérant, Intelligent.*

Cette seconde description reprend les mêmes adjectifs, et pourtant, à cause de l'ordre dans lequel ils sont présentés, elle est bien plus défavorable que la première. Le mot « Rusé » n'était que légèrement négatif quand il venait à la suite des mots « Intelligent » et « Persévérant », car nous supposions encore (sans raison valable) que les intentions du candidat étaient bonnes. Mais quand on est déjà « sans scrupule », être « rusé » devient une circonstance aggravante. Quant à l'intelligence et à la persévérance, elles n'ont plus rien de positif : au contraire, elles rendent une personne mauvaise plus dangereuse encore.

Cette expérience illustre l'*excès de cohérence* : nous nous forgeons très vite des impressions cohérentes, et nous sommes trop lents et trop réticents à les modifier[10]. Dans notre exemple, il a suffi de très peu d'éléments pour que nous nous fassions une opinion positive du candidat. Le biais de confirmation – cette même tendance qui nous conduit, quand nous avons un

préjugement, à négliger les éléments qui pourraient l'infirmer – nous fait donner moins d'importance qu'il ne le faudrait aux éléments qui viennent ensuite. (Une autre expression pour décrire ce phénomène est l'*effet de halo*, parce que le candidat bénéficie du « halo » positif de la première impression. Nous verrons au chapitre 24 que l'effet de halo est un problème sérieux dans les décisions de recrutement.)

Une étude dans un domaine bien différent illustre elle aussi l'excès de cohérence. Aux États-Unis, les chaînes de restaurants ont l'obligation d'afficher des informations nutritionnelles, afin que les consommateurs n'ignorent rien du nombre de calories contenues dans les hamburgers ou les salades qui leur sont proposés. Ces informations modifient-elles les choix des consommateurs ? Les études sur la question sont partagées. L'une de ces études a cependant démontré un résultat étrange : les indications caloriques ont un effet plus prononcé sur les comportements des clients lorsqu'elles sont placées *à gauche* du nom des plats que lorsqu'elles sont placées à leur droite[11]. Quand les calories sont indiquées à gauche, c'est l'information que les clients lisent en premier : manifestement, ils pensent « trop de calories ! » ou « pas tant de calories que ça ! », avant même de lire le nom du plat. Leur réaction initiale, positive ou négative, colore la suite de leur jugement. En revanche, lorsque les clients lisent d'abord le nom du plat, il semble qu'ils pensent « irrésistible » ou « pas très appétissant » avant de lire les informations nutritionnelles. Là encore, leur réaction première affecte fortement leurs choix. Cette hypothèse est étayée par une expérience complémentaire menée en Israël, et où l'on présente aux sujets des menus écrits en hébreu, donc lus de droite à gauche. Dans cette langue, c'est bien quand l'indication calorique est placée à droite du nom du plat qu'elle a un impact plus important.

De manière générale, nous avons tendance à nous forger hâtivement des convictions, puis à nous y tenir. Nous pensons que nos convictions reposent sur les faits, mais, au moins dans une certaine mesure, c'est le contraire : le

choix des faits dont nous tenons compte et l'interprétation que nous en faisons sont déformés pour conforter notre jugement premier. Ce mécanisme nous permet de préserver la cohérence de l'histoire qui s'est construite dans notre esprit.

Bien sûr, si notre conviction de départ est correcte, cette quête de cohérence n'est pas dommageable. Mais si le jugement initial est erroné, la tendance à s'y accrocher en négligeant les faits qui pourraient l'infirmer est une source d'erreurs. Qui plus est, il est difficile d'échapper à ce mécanisme mental, parce qu'il est presque impossible de se forcer à ignorer ou à oublier les informations auxquelles nous avons été exposés et les conclusions que nous en avons déjà tirées. Dans les salles d'audience américaines, il arrive que les juges demandent aux jurés d'ignorer un élément qui leur a été présenté, parce que cet élément est jugé juridiquement irrecevable. Hélas, il n'est pas très réaliste d'espérer que les jurés se conforment à cette instruction (sauf peut-être durant leurs délibérations, où les arguments fondés expressément sur un élément irrecevable pourront être rejetés).

## *Du biais psychologique au biais statistique… et au bruit*

Nous avons brièvement présenté trois types de biais, qui agissent de trois façons différentes. Les *biais de substitution* produisent une pondération erronée des données du problème considéré. Les *biais de conclusion* nous conduisent soit à court-circuiter l'évaluation de ces données, soit à les traiter d'une manière déformée. Enfin, l'*excès de cohérence* accentue l'effet d'une première impression et limite l'impact des informations qui contredisent cette impression.

Les trois types de biais peuvent, bien sûr, produire du biais statistique. Mais ils peuvent aussi produire du bruit.

Commençons par les biais de substitution. Dans l'exemple de Bill, la plupart des gens évaluent la probabilité que Bill exerce la profession de comptable en évaluant sa ressemblance avec un stéréotype. Si tous les sujets de l'expérience font la même erreur et partagent le même stéréotype du comptable, le résultat est un biais partagé, et il n'y a pas de bruit. Mais la substitution d'une question à une autre n'engendre pas nécessairement une parfaite unanimité. Quand on remplace la question « Le changement climatique est-il une réalité ? » par « Est-ce que je fais confiance aux gens qui parlent du changement climatique ? », il est évident que la conclusion variera d'une personne à l'autre, en fonction du milieu social de cette personne, de ses sources d'information privilégiées, de sa couleur politique, et ainsi de suite. Le biais psychologique – la substitution d'une question à une autre – est le même que dans l'exemple de Bill ; mais ici, il crée de la variabilité dans les jugements, c'est-à-dire du bruit.

La substitution peut aussi être source de bruit occasionnel. Si l'on répond à une question sur sa satisfaction dans la vie en consultant son humeur du moment, la même personne donnera forcément des réponses variables d'un moment à un autre. Une matinée joyeuse peut être suivie d'une triste après-midi, et ces changements d'humeur vous conduiront à évaluer différemment votre satisfaction dans la vie, en fonction du moment où un sondeur vous téléphone pour vous poser la question. De la même manière, beaucoup des exemples de bruit occasionnel que nous avons mentionnés au chapitre 7 sont déclenchés par des biais psychologiques qui font substituer une question à une autre.

Les biais de conclusion, et notamment les préjugements, produisent aussi des erreurs qui sont tantôt du biais, tantôt du bruit. Revenons à un exemple évoqué dans l'introduction de cet ouvrage : les scandaleuses disparités entre juges dans la proportion des demandes d'asile auxquelles ils

accèdent. Quand un juge accorde l'asile à 5 % des demandeurs alors qu'un autre juge, dans le même tribunal, en admet 88 %, nous pouvons avoir la quasi-certitude que ces juges sont biaisés dans des directions très différentes.

Cet exemple illustre un principe général. Un système peut, bien sûr, être globalement affecté par un biais statistique, si la plupart ou la totalité des juges sont biaisés dans le même sens. Mais quand il existe des différences individuelles dans les biais des juges, ces différences sont source de bruit systémique – dans des proportions qui peuvent être considérables.

Qu'en est-il, enfin, de l'excès de cohérence ? Comme les deux premiers types de biais, celui-ci pourra être source soit de biais, soit de bruit. Selon que l'ordre dans lequel les informations sont présentées et le sens qui leur est attribué sont ou non identiques pour tous les juges (ou pour la plupart d'entre eux), le résultat sera bien différent.

Considérons, par exemple, un candidat physiquement séduisant qui suscite une première impression positive chez la plupart des recruteurs. Le poids de cette première impression et le halo positif qu'elle crée se traduiront par une erreur partagée (du moins si l'on suppose que l'apparence physique n'est pas un critère pertinent de recrutement pour le poste visé). Le résultat, ici, est un biais statistique.

Mais si, dans cet exemple, la première impression est la même pour tout le monde, ce n'est pas une règle générale. Beaucoup de décisions complexes reposent, au contraire, sur la compilation d'informations qui peuvent se présenter dans n'importe quel ordre. Pour les experts en sinistres du chapitre 2, par exemple, l'ordre dans lequel arrivent les informations sur un sinistre varie d'un expert à l'autre et d'un sinistre à l'autre. Les premières impressions varient donc de manière aléatoire ; et l'excès de cohérence traduit ces variations des premières impressions en distorsions dans les jugements finaux. Le résultat n'est pas un biais statistique, mais du bruit systémique.

En résumé, les biais psychologiques – de substitution, de conclusion ou d'excès de cohérence – sont des mécanismes mentaux universels. Ils produisent donc souvent des erreurs partagées. Mais quand il existe de grandes différences dans les biais des individus (notamment des biais de conclusion différents), ou quand l'effet des biais dépend d'éléments déclencheurs différents dans le contexte d'un jugement donné, les biais psychologiques sont source de bruit.

Comme, quoi qu'il en soit, le biais statistique et le bruit sont deux composantes de l'erreur, tout ce qui peut contribuer à réduire les biais psychologiques est bon pour la qualité des jugements. Nous reviendrons dans la partie V sur la lutte contre les biais. Mais avant cela, poursuivons notre exploration du processus de jugement.

### À propos des heuristiques, des biais et du bruit

« Nous savons que nous sommes sujets à des biais psychologiques, mais nous devrions résister à la tentation de mettre toutes les erreurs sur le compte de "biais" non identifiés. »

« Quand nous substituons une question facile à la question difficile qui nous est posée, nous faisons inévitablement des erreurs. Par exemple, quand nous répondons en fonction de la ressemblance alors que la question porte sur la probabilité, nous négligeons le taux de base. »

« Nos préjugements et nos autres biais de conclusion ont déformé les informations dont nous disposions de manière à conforter notre opinion initiale. »

« Nous nous sommes forgé rapidement une impression, et nous nous y accrochons alors que surviennent des informations contraires : nous sommes victimes de l'excès de cohérence. »

« Les biais psychologiques engendrent du biais statistique, à condition que beaucoup de gens partagent les mêmes biais. Souvent, ce n'est pas le cas ; et les biais psychologiques ne créent alors pas du biais statistique, mais du bruit systémique. »

CHAPITRE 14

# Équivalences

Levez les yeux et regardez le ciel. Quelle est la probabilité qu'il pleuve dans les deux heures ?

Vous n'avez sans doute eu aucun mal à répondre à cette question. Le jugement que vous avez produit – par exemple, « très probable » – ne vous a demandé aucun effort. D'une manière ou d'une autre, l'évaluation que vous avez faite de la couleur du ciel a été convertie en un jugement de probabilité.

Ce que vous venez de faire est un exemple élémentaire du mécanisme mental qui nous permet d'*établir des équivalences*. Nous avons défini le jugement comme le fait d'attribuer à une impression (ou à un aspect d'une impression) une certaine valeur sur une certaine échelle. La recherche d'équivalences est indispensable à cette opération. Quand vous répondez à la question « Sur une échelle de 1 à 10, quelle est votre humeur ? », ou « Sur une échelle de 1 à 5 étoiles, quelle est votre satisfaction après votre visite dans ce magasin ? », vous établissez une équivalence entre une valeur sur l'échelle proposée et votre humeur ou votre expérience.

# *Équivalence et cohérence*

Nous allons maintenant retrouver Bill, dont vous avez fait la connaissance au chapitre précédent. Souvenez-vous : « Bill a 33 ans. Il est intelligent mais sans imagination, compulsif et, de façon générale, assez amorphe. À l'école, il était fort en mathématiques mais faible dans les autres matières, en particulier littéraires. » Nous vous avons demandé d'estimer la probabilité que Bill exerce diverses professions et qu'il s'adonne à différents loisirs, et nous avons vu que vous vous êtes acquitté de cette tâche en substituant des jugements de ressemblance à des jugements de probabilité. Par exemple, vous ne vous êtes pas vraiment demandé quelle était la probabilité que Bill soit comptable, mais plutôt dans quelle mesure il ressemblait au stéréotype de cette profession. Nous allons maintenant nous demander *comment* vous avez produit ce jugement de ressemblance.

Il n'est pas difficile d'évaluer le degré de correspondance entre la description qui vous a été donnée de Bill et des stéréotypes de métiers et de hobbies. Manifestement, Bill ressemble moins au cliché du joueur de jazz qu'à celui du comptable, et encore moins à celui du surfeur. La facilité avec laquelle nous produisons ces jugements illustre l'extraordinaire flexibilité avec laquelle nous trouvons des équivalences, en particulier dans les jugements sur les personnes.

Il n'y a quasiment pas de limite aux questions auxquelles vous pourriez répondre au sujet de Bill. Par exemple, que penseriez-vous de la perspective de vous retrouver seul avec lui sur une île déserte ? Une réponse immédiate et intuitive à cette question vous est sans doute déjà venue à l'esprit, sur la base du peu d'informations dont vous disposez. Or nous avons du nouveau : il se trouve que Bill, que nous avons la chance de bien connaître, est un explorateur endurci qui possède des aptitudes extraordinaires à la survie en environnement hostile ! Si cette information vous surprend (et nous soupçonnons que c'est le cas), c'est qu'elle vous pose un sérieux problème

de cohérence. La surprise est forte parce que l'image d'un aventurier aguerri, boussole dans la poche et couteau fixé au ceinturon, est radicalement incompatible avec l'idée que vous vous étiez faite de Bill.

Imaginez maintenant que cet aspect inattendu de la personnalité de Bill ait fait partie de la première description que vous avez lue. L'image de Bill que vous vous seriez construite aurait été tout autre : peut-être auriez-vous imaginé le genre de personne qui s'étiole dans la grisaille des villes, mais dont la vraie personnalité se révèle dans les grands espaces. L'impression générale de Bill aurait été moins cohérente, et vous auriez peut-être trouvé plus difficile d'évaluer son degré de correspondance avec les métiers ou les hobbies qui vous étaient proposés. Mais vous n'auriez pas éprouvé le violent sentiment de dissonance que vous a occasionné la révélation tardive de ses talents cachés.

Comme l'illustre cet exemple, en présence d'indications contradictoires, il devient plus difficile de parvenir au sentiment de cohérence nécessaire à la formulation d'un jugement satisfaisant. La présence d'indications contradictoires caractérise les jugements complexes, dans lesquels on peut par conséquent s'attendre à trouver beaucoup de bruit. Le problème Gambardi était un jugement de ce type : certaines des indications sur Gambardi étaient positives, d'autres négatives. Nous reviendrons sur les problèmes que posent les jugements complexes au chapitre 16. Dans l'immédiat, nous allons nous tourner vers des jugements relativement simples, et notamment ceux qui utilisent des *échelles d'intensité*.

## *Les équivalences d'intensité*

Certaines des échelles que nous utilisons pour formuler des jugements sont qualitatives : par exemple, les jugements qui utilisent des catégories (comme les hobbies possibles de Bill, ou les diagnostics possibles pour un

patient). On les reconnaît au fait que leurs valeurs ne sont pas ordonnées. Sur l'échelle des couleurs, par exemple, « rouge » n'est ni « plus » ni « moins » que « bleu ». Sur celle des métiers, il n'y a pas d'ordre entre « comptable » et « architecte ».

Mais un grand nombre de jugements font appel à des échelles d'intensité quantitatives. C'est le cas des mesures physiques comme la taille, le poids, la luminosité, la température ou le volume sonore ; des mesures de coût ou de valeur ; des jugements de fréquence ou de probabilité. C'est le cas aussi quand les échelles d'intensité sont plus abstraites. La confiance, la force, la séduction, la colère, la peur, l'immoralité des crimes ou la sévérité des peines sont toutes des échelles d'intensité.

Ce qui caractérise ces échelles d'intensité quantitatives est qu'on peut répondre à la question : « Lequel des deux est plus… ? », qu'il s'agisse d'être plus grand, chaud, cher, séduisant, immoral ou sévère. On peut dire qu'une flagellation est une peine plus sévère qu'une tape sur le poignet, ou trouver que les mérites littéraires de *Hamlet* surpassent ceux d'*En attendant Godot*, exactement de la même manière qu'on peut dire que le soleil est plus brillant que la lune, qu'un éléphant pèse plus lourd qu'un hamster et que la température moyenne à Miami est plus élevée qu'à Toronto.

L'intérêt de ces échelles d'intensité tient à une remarquable aptitude que nous avons tous : nous pouvons établir intuitivement une équivalence entre des jugements d'intensité sur des dimensions qui n'ont aucun rapport entre elles. Il nous suffit pour cela de calquer une échelle d'intensité sur une autre[1]. Vous pouvez aisément trouver une équivalence entre le talent de différents chanteurs et la hauteur des bâtiments de votre ville : si vous trouvez que Bob Dylan est un génie, par exemple, votre niveau de ferveur pour lui correspond au plus haut édifice de la ville. Vous pourriez faire correspondre le niveau actuel de virulence du débat politique dans votre pays avec une température estivale dans sa capitale : si les tempéraments

sont particulièrement échauffés, vous choisirez une température caniculaire de 38 °C à Paris. Et si l'on vous demandait d'exprimer votre avis sur un restaurant en le comparant à la longueur d'un roman plutôt qu'en lui attribuant des étoiles, cela vous semblerait certainement bizarre, mais pas du tout impossible : votre restaurant favori pourrait être *Guerre et paix*. Dans chacun de ces exemples, ce que vous voulez dire est – étrangement – très clair.

Dans la conversation ordinaire, l'éventail des valeurs sur une échelle est fonction du contexte. La phrase : « Elle a mis beaucoup d'argent de côté » a un sens différent si l'on célèbre le départ à la retraite d'une grande banquière d'affaires ou si l'on félicite une adolescente qui fait régulièrement du baby-sitting. Plus généralement, la signification de mots comme « beaucoup », « grand » ou « petit » dépend entièrement du cadre de référence. C'est parce que nous partageons un cadre de référence que nous pouvons comprendre une phrase comme : « La *grosse* souris grimpe sur la trompe du *petit* éléphant. »

## *Les biais dans les prédictions par équivalence*

La devinette suivante illustre à la fois la puissance des équivalences entre échelles d'intensité et une erreur de jugement prévisible qui en résulte [2].

> *Julie est étudiante dans une université américaine. Lisez ces informations à son sujet, puis essayez de deviner la moyenne générale de ses notes à l'université (son GPA ou* grade point average, *sur l'échelle standard américaine de 0,0 à 4,0).*
> *Julie lisait couramment à l'âge de 4 ans.*

*Quel est son GPA ?*

Si vous êtes familier du système de notation scolaire américain, un chiffre vous est certainement venu à l'esprit très vite, probablement proche de 3,7 ou 3,8. La manière dont il vous est venu à l'esprit dans le cas de Julie fait appel à un jugement par équivalence, analogue à ceux que nous venons de décrire.

Pour commencer, vous avez évalué dans quelle mesure Julie était une lectrice précoce. Cela n'était pas difficile : elle a su lire à un âge inhabituellement jeune, qui la place dans une certaine catégorie sur l'échelle de précocité. Si vous deviez décrire cette échelle, vous définiriez probablement sa catégorie la plus élevée comme « extraordinairement précoce », et vous noteriez que Julie n'y rentre pas (certains enfants savent lire avant l'âge de 2 ans). Julie appartient probablement à la catégorie suivante, le groupe des enfants « inhabituellement mais pas extraordinairement précoces ».

Dans une seconde étape, vous avez cherché une estimation de GPA qui corresponde avec votre évaluation de la précocité de Julie. Sans en avoir forcément conscience, vous avez gravité vers une valeur de GPA qui puisse aussi être qualifiée d'« inhabituelle mais pas extraordinaire ». Ainsi, et bien sûr sous réserve que l'échelle des GPA vous soit familière, une *prédiction par équivalence* vous est venue à l'esprit, apparemment surgie de nulle part, quand vous avez lu la description de Julie.

Naturellement, vous n'avez pas suivi ces deux étapes (évaluation et recherche d'équivalence) de manière méthodique et délibérée, en faisant les calculs statistiques nécessaires : cela vous aurait pris un temps considérable. C'est la pensée rapide du Système 1 qui vous a permis d'arriver à un jugement vite et sans effort.

La séquence complexe d'opérations mentales que nous venons de proposer pour expliquer ce mécanisme n'est pas directement observable. Si

nous décrivons ce processus de prédiction par équivalence avec une précision inhabituelle en psychologie, c'est parce que les éléments qui l'attestent sont inhabituellement concluants. De fait, sur la base de nombreuses expériences similaires, nous pouvons être certains que les réponses aux deux questions suivantes, quand elles sont posées à des groupes différents de personnes, seront strictement identiques[3] :

- ❑ *Quel pourcentage des enfants nés la même année que Julie ont lu à un âge plus précoce qu'elle ?*
- ❑ *Quel pourcentage d'élèves de la promotion de Julie ont un GPA supérieur au sien ?*

Il y a une énorme différence entre ces deux questions. Pour répondre à la première, il suffit de mettre en perspective l'information qui vous a été fournie sur Julie. La seconde est bien plus difficile, puisqu'elle exige que vous fassiez une prédiction sur un sujet éloigné de l'information dont vous disposez.

Il est pourtant intuitivement très tentant de répondre à la seconde question en se servant de la première. Vous reconnaissez ici la distinction que nous avons introduite en présentant l'illusion de validité (distinction dont nous avons souligné qu'elle n'est naturelle pour personne). La première question nécessite d'évaluer l'« intensité » de l'*information* dont nous disposons. La seconde nous interroge sur l'« intensité » d'une *prédiction*. La distinction entre information et prédiction était déjà difficile quand vous confondiez votre évaluation des qualités présentes d'un candidat avec une prévision de ses chances de succès futur. Elle demeure difficile dans cet exemple.

Vous reconnaissez aussi dans cette prévision intuitive du GPA de Julie un exemple du mécanisme psychologique décrit au chapitre précédent : la substitution d'une question facile à une question difficile. Votre Système 1

simplifie une question difficile de prédiction en répondant à une question beaucoup plus facile : dans quelle mesure le fait que Julie ait su lire à l'âge de 4 ans est-il remarquable ? Il suffit alors d'une étape supplémentaire de recherche d'équivalence pour que l'âge d'apprentissage de la lecture, mesuré en années, soit intuitivement traduit en un score de GPA, mesuré en points.

Bien sûr, la substitution n'a lieu que si les informations fournies semblent pertinentes. Si tout ce qu'on vous disait de Julie enfant est qu'elle courait vite ou qu'elle était une ballerine médiocre, vous ne sauriez que faire de ces indications. Mais dès lors qu'une information peut être interprétée comme une indication plausible d'intelligence, elle servira probablement de base à un jugement par substitution.

La substitution d'une question à une autre est nécessairement source d'erreurs, dès lors que les véritables réponses aux deux questions sont différentes. Même si le Système 1 trouve plausible d'utiliser l'âge de maîtrise de la lecture pour prédire la note moyenne à l'université, cette substitution est manifestement absurde. Il suffit pour vous en convaincre de songer à quelques-uns des événements qui auraient pu survenir dans la vie de Julie entre ses 4 ans et son entrée à l'université. Julie aurait pu avoir un terrible accident. Ses parents auraient pu vivre un divorce traumatisant. Elle aurait pu tomber enceinte. Tous ces événements, et une infinité d'autres, auraient pu affecter ses résultats scolaires.

Une prédiction par équivalence ne se justifierait que si la précocité de lecture et la note moyenne à l'université étaient parfaitement corrélées, ce qui n'est évidemment pas le cas. Néanmoins, il serait dommage d'ignorer complètement l'information dont vous disposez sur l'âge auquel lisait Julie, car cette information n'est pas entièrement dénuée de pertinence. La prédiction optimale doit donc se situer quelque part entre deux extrêmes : la prédiction par équivalence, qui suppose implicitement que vous avez une

connaissance parfaite des talents de Julie, et celle qui fait entièrement abstraction de toute information à son sujet.

Justement, si vous ne saviez rien de Julie, à part qu'elle est étudiante à l'université, comment vous y prendriez-vous pour estimer sa moyenne ? Cette question fait appel à ce que nous avons appelé la vision externe : la vision que vous adoptez sur un cas particulier quand vous n'avez aucune information spécifique sur ce cas, sauf une catégorie à laquelle il appartient. À l'évidence, si on ne vous donnait aucune information sur Julie, votre prédiction de son GPA serait certainement le GPA moyen des étudiants à l'université, soit – pour un lecteur familier du système de notation nord-américain – une note aux alentours de 3,2. Ce chiffre, dans notre exemple, est la *prédiction par vision externe*.

La meilleure estimation possible du GPA de Julie doit donc se situer entre la prédiction par équivalence et la prédiction par vision externe : quelque part entre 3,2 et 3,8. La réponse précise dépend de la valeur prédictive des informations qui servent de base à l'équivalence. Plus vous pensez que l'âge de maîtrise de la lecture est un bon prédicteur du GPA, plus votre estimation devra être proche de 3,8. Ici, on l'a vu, la valeur prédictive de cette information est certainement faible, et la prédiction la plus raisonnable doit donc être plus proche de 3,2. Il existe une technique relativement simple pour affiner ce calcul, et de manière générale pour corriger l'erreur inhérente aux prédictions par équivalence : nous la présentons en détail dans l'annexe C.

En dépit de l'absurdité statistique des résultats auxquels conduisent les prédictions par équivalence, il est très difficile d'y résister. Il semble intuitivement très naturel, pour formuler une prévision, de s'appuyer sur les données disponibles et de rechercher des équivalences. Les responsables de ventes supposent souvent que le meilleur vendeur de l'année passée va continuer à surperformer. Les recruteurs qui rencontrent en entretien un candidat exceptionnellement talentueux s'imaginent l'ascension de la

nouvelle recrue jusqu'au sommet de l'organisation. Et quand le dernier film d'un réalisateur a été un grand succès, les producteurs escomptent bien que son film suivant en sera un à son tour.

Les attentes que font naître ces prédictions par équivalence ont toutes les chances d'être déçues. Inversement, si vous faites des prédictions par équivalence quand tout va mal, celles-ci risquent fort d'être excessivement défavorables. Les prédictions par équivalence sont trop extrêmes, qu'elles soient optimistes ou pessimistes. (Le terme technique pour ces prédictions intuitives est qu'elles sont *non régressives*, parce qu'elles omettent de prendre en compte un phénomène statistique appelé la *régression vers la moyenne*.)

Il faut noter, cependant, que la substitution et la recherche d'équivalence ne régissent pas toutes les prédictions. Pour reprendre le langage des deux systèmes, le Système 1, intuitif, propose des réponses associatives et rapides aux questions qui se présentent à nous, mais avant de devenir des croyances ou des convictions, ces intuitions doivent être confirmées par le Système 2, réfléchi. Il arrive donc que des prédictions par équivalence soient rejetées au profit de réponses plus complexes. En particulier, on est moins prompt à formuler des prédictions par équivalence quand les données disponibles sont défavorables. Si l'on vous avait dit que Julie n'avait su lire qu'à un âge relativement avancé, vous hésiteriez sans doute avant d'en conclure que sa moyenne à l'université doit être médiocre. En tout cas, vous hésiteriez plus longtemps que vous n'avez hésité à proposer une équivalence quand la donnée était favorable. (Cette asymétrie entre prédictions favorables et défavorables disparaît quand la quantité d'informations fournies s'accroît.)

Pour corriger les prédictions intuitives, quelles qu'elles soient, une solution s'impose : la vision externe. Dans l'exemple de Michael Gambardi, par exemple, nous avons recommandé d'ancrer votre jugement de la probabilité de réussite de Michael sur le taux de base pertinent (le taux de

« survie » à deux ans d'un P-DG moyen). S'agissant de prédictions quantitatives, comme le GPA de Julie, adopter la vision externe consiste à ancrer votre prédiction sur la moyenne de la catégorie à laquelle appartient le cas étudié. Sauf dans les problèmes très simples, où les informations disponibles permettent de faire une prédiction avec une confiance totale, la vision externe est indispensable. Dès lors qu'un jugement sérieux est nécessaire, elle doit faire partie de la solution.

## *Du bruit dans les équivalences : les limites du jugement absolu*

Quand nous établissons des équivalences entre des échelles d'intensité, la précision de nos jugements a une limite : celle de notre capacité à distinguer des catégories sur chaque échelle. Des mots comme *grand* ou *riche* peuvent désigner une grande variété de valeurs différentes de la taille ou de la fortune. Cette ambiguïté est une importante source de bruit.

Ainsi, nous pouvons sans doute qualifier la banquière d'investissement qui prend sa retraite de *riche*, mais que signifie le mot, exactement ? D'autres épithètes auraient pu être retenues : *aisée*, *fortunée*, *super-riche*, etc. Si l'on vous proposait une description précise de la fortune de différents individus et qu'on vous demandait de les ranger dans des catégories définies par un adjectif différent, combien de catégories distinctes pourriez-vous créer – sans recourir à des comparaisons détaillées entre les individus ?

La question générique que pose cet exemple est celle du nombre de catégories que nous pouvons distinguer sur une échelle d'intensité. La réponse figure dans le titre d'un article de 1956 devenu un classique de la psychologie : « Le chiffre magique de sept, plus ou moins deux[4]. » Au-delà

de cette limite, les gens ont tendance à faire des erreurs, par exemple à ranger A dans une catégorie plus élevée que B, tandis qu'ils donnent une note plus élevée à B qu'à A quand ils les comparent directement l'un avec l'autre.

Pour illustrer ce processus de catégorisation, imaginez une série de lignes mesurant toutes entre 2 et 4 centimètres, mais qui sont de quatre longueurs différentes, également espacées entre les deux extrêmes : les lignes mesurent donc 2, 2,67, 3,33 ou 4 centimètres. On vous les présente une par une. Chaque fois, vous devez indiquer un chiffre de 1 à 4 : 1 quand la ligne appartient à la catégorie des lignes les plus courtes ; 2 pour la deuxième catégorie, etc. L'exercice est facile, et vous ne commettez aucune erreur. Supposez maintenant que l'on vous montre des lignes de cinq longueurs différentes, à catégoriser de 1 à 5. Facile encore. Quand allez-vous commencer à vous tromper ? Autour du chiffre magique de 7. Étonnamment, ce résultat ne dépend pas de l'espacement entre les lignes : si leurs longueurs allaient de 2 à 6 centimètres et non de 2 à 4 centimètres, c'est quand même au-delà de 7 longueurs que vous vous mettriez à faire des erreurs. On obtient à peu près le même résultat avec d'autres types de stimuli, comme des sons ou des lumières dont on fait varier l'intensité. Il y a une réelle limite à notre capacité d'assigner à des stimuli différents des labels différents sur une dimension donnée, et cette limite se situe autour de 7.

Cette limite à notre capacité de discrimination est essentielle. En effet, notre aptitude à établir des équivalences entre des échelles d'intensité différentes ne peut pas être supérieure à notre capacité à attribuer des valeurs sur ces mêmes échelles. Autrement dit, si nous nous trompons dans l'utilisation de chaque échelle, nous avons de fortes chances de nous tromper quand nous en associons deux. Les équivalences entre échelles d'intensité sont un outil polyvalent de la pensée rapide du Système 1, qu'on

retrouve au cœur d’un grand nombre de jugements intuitifs. Mais elles sont assez rudimentaires.

Le chiffre magique de 7 n’est pas une limite absolue. On peut s’exercer à faire des distinctions plus fines en hiérarchisant des catégories et des sous-catégories. Ainsi, on peut certainement distinguer plusieurs sous-catégories de richesse au sein de la catégorie des multimillionnaires. De même, les juges peuvent non seulement hiérarchiser des infractions par ordre de gravité, mais aussi distinguer, au sein de chaque infraction, des circonstances ou des affaires présentant des degrés de gravité différents. Toutefois, pour que ce type de catégorisation soit possible, les catégories doivent être préalablement définies et leurs frontières doivent être nettes. Pendant que vous êtes en train de chercher les catégories qui correspondent à la longueur de différentes lignes, vous ne pouvez pas décider de séparer les lignes les plus longues des lignes les plus courtes et de les considérer comme deux supercatégories distinctes. Lorsqu’on utilise le mode de pensée rapide, la catégorisation ne peut pas être contrôlée à volonté.

## *Les vertus du jugement comparatif*

Il existe un moyen de surmonter la résolution limitée des échelles qui utilisent des épithètes ou des descriptions comme labels pour leurs catégories : remplacer ces labels par des comparaisons. Notre capacité de comparer des cas entre eux est bien supérieure à notre capacité de les placer sur une échelle.

Demandez-vous ce que vous feriez si l’on vous demandait d’évaluer un grand nombre de restaurants, ou de chanteurs, sur une échelle allant de 1 à 20 points. Une échelle allant d’une à cinq étoiles est facile à manier, mais avec une échelle allant de 1 à 20, il est difficile de conserver une fiabilité

parfaite. Joe’s Pizza vaut trois étoiles, mais mérite-t-il un 11 ou plutôt un 12 ?

La solution à ce problème est simple, bien qu’un peu laborieuse. Il suffit de noter d’abord les restaurants ou les chanteurs en utilisant les cinq étoiles pour les répartir en autant de catégories. Vous pourrez ensuite les hiérarchiser au sein de chaque catégorie. Il y aura peut-être quelques *ex aequo*, mais vous savez sans doute si vous préférez Joe’s Pizza à Fred’s Burgers, ou Taylor Swift à Bob Dylan, même si vous leur avez donné le même nombre d’étoiles. Par souci de simplicité, vous pourriez distinguer quatre niveaux à l’intérieur de chaque catégorie. L’exercice n’est pas difficile : même entre les chanteurs que vous détestez le plus, vous seriez sans doute capable de distinguer des niveaux différents d’aversion.

Cet exercice repose sur un principe psychologique simple. Les comparaisons explicites entre objets permettent des jugements plus fins que la notation d’objets évalués un à un. Si vous reveniez à l’exercice consistant à comparer des lignes de longueurs différentes, vous feriez le même constat : il vous est bien plus facile de comparer les longueurs de lignes qui vous sont montrées immédiatement l’une après l’autre que d’estimer leurs longueurs. Vous seriez même encore plus précis si les lignes à comparer vous étaient toutes présentées en même temps.

La supériorité des jugements comparatifs se vérifie dans un grand nombre de domaines. Si vous avez une vague idée de la fortune d’un certain nombre d’individus, vous aurez intérêt, pour en avoir une idée plus précise, à les comparer entre eux plutôt qu’à évaluer leur richesse individuellement. Si vous notez des dissertations, vous serez plus précis en classant toutes les copies de la meilleure à la moins bonne qu’en les lisant et en les notant une par une. Les jugements comparatifs ou relatifs sont plus fins que les jugements catégoriques ou absolus, en ce sens qu’ils permettent plus de différenciation. Comme le montrent ces exemples, ils demandent aussi, bien sûr, plus d’efforts.

Comment faire pour conserver les avantages des jugements comparatifs quand on doit évaluer des objets individuellement ? On peut pour cela s'efforcer d'utiliser des échelles explicitement comparatives. Dans certains contextes, notamment universitaires, on demande souvent aux auteurs d'une lettre de recommandation de situer la candidate qu'ils recommandent par rapport à une population donnée : de dire, par exemple, si elle appartient aux « meilleurs 5 % » ou aux « meilleurs 20 % » des « étudiants auxquels vous avez enseigné » ou des « diplômés ayant le même nombre d'années d'expérience ».

Ces évaluations ne peuvent bien sûr pas être prises pour argent comptant, puisqu'il est généralement impossible de s'assurer de la rigueur avec laquelle les auteurs des recommandations utilisent l'échelle. Toutefois, dans certains contextes, il devient possible de le vérifier, par exemple quand des cadres évaluent un grand nombre d'employés, ou des analystes un grand nombre d'investissements. Si l'un d'entre eux a tendance à ranger 90 % des cas dans la catégorie des « meilleurs 20 % », on pourra rapidement s'en rendre compte et l'inviter à recalibrer son jugement. L'utilisation des jugements comparatifs est donc un remède possible au bruit : nous y reviendrons dans la partie V.

---

De nombreuses tâches de jugement consistent à ranger des cas individuels dans des catégories, comme les degrés d'une échelle d'assentiment (de « tout à fait d'accord » à « pas du tout d'accord ») ou un ensemble ordonné d'épithètes (un événement « probable » ou « extrêmement probable »). Nous n'éprouvons guère de difficulté à réaliser cette opération, car nous trouvons aisément des équivalences entre des échelles d'intensité différentes. Toutefois, si les échelles sont imprécises et ne permettent pas des jugements suffisamment fins, les équivalences ainsi établies entre cas et catégories seront sujettes au bruit. Une procédure qui rend les jugements explicitement comparatifs peut aider à réduire ce bruit[5]. Dans le chapitre

suivant, nous montrerons que l’emploi d’échelles mal choisies peut, au contraire, l’aggraver.

### À propos des équivalences

« Nous disons tous les deux que ce film est très bon, mais il me semble que vous l’avez beaucoup moins aimé que moi. Nous utilisons les mêmes mots, mais comprenons-nous bien l’échelle de la même manière ? »

« La Saison 1 de la série était exceptionnelle ; et nous pensions que la Saison 2 le serait aussi. C’était une prédiction par équivalence : nous nous sommes trompés. »

« Il est difficile de rester cohérent quand on note des dissertations. Faudrait-il plutôt essayer de les classer ? »

CHAPITRE 15

# Échelles

Imaginez que vous soyez juré dans un procès au civil. Vous avez pris connaissance des faits résumés ci-dessous, et on attend de vous plusieurs jugements sur l'affaire.

*L'AFFAIRE JOAN GLOVER VS GENERAL ASSISTANCE*

*Joan Glover, une enfant de 6 ans, a ingéré un grand nombre de cachets d'Allerfree, un médicament antiallergique en vente libre. Elle a dû subir une longue hospitalisation. Comme l'intoxication a affaibli son système respiratoire, elle sera, pour le restant de ses jours, plus exposée au risque de maladies respiratoires comme l'asthme et l'emphysème. Le bouchon du flacon contenant les cachets d'Allerfree était équipé d'une sécurité enfant mal conçue.*

*Le fabricant d'Allerfree est General Assistance, une grande entreprise de médicaments vendus sans ordonnance, dont les profits annuels sont entre 100 et 200 millions de dollars. Une réglementation fédérale exige des bouchons avec sécurité enfant sur tous les flacons de médicaments. General Assistance a ignoré systématiquement l'esprit de cette réglementation en équipant ses flacons d'un bouchon qui a un taux de défaillance nettement plus élevé que ceux des autres fabricants. Un*

*document interne de l'entreprise déclare que « cette réglementation fédérale stupide et inutile nous fait perdre de l'argent » et affirme que General Assistance ne court que très peu de risques d'être sanctionnée. Le document ajoute qu'en tout état de cause « les pénalités pour manquement à la réglementation sont extrêmement légères ; en gros, on nous demandera d'améliorer la sécurité des bouchons dans l'avenir ». Bien qu'un agent de la Food and Drug Administration ait officiellement signalé le problème à General Assistance, l'entreprise n'a pris aucune mesure pour y remédier.*

Nous vous demanderons ensuite de produire trois jugements et de noter vos réponses.

**Indignation :**
Laquelle de ces expressions exprime le mieux votre opinion sur les actions de l'entreprise mise en cause ? (Entourez SVP le chiffre qui *exprime* le mieux votre opinion.)

| Pleinement acceptables | | Contestables | | Choquantes | | Absolument scandaleuses |
|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 |

**Intention de sanctionner :**
En plus de payer à la victime des dommages et intérêts compensatoires, comment l'entreprise mise en accusation devrait-elle être punie ? (Entourez SVP d'un cercle le chiffre qui exprime le mieux votre opinion sur le niveau approprié de la sanction.)

| Aucune sanction | | Sanction modérée | | Sanction sévère | | Sanction extrêmement sévère |
|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 |

**Dommages :**
En plus des dommages et intérêts compensatoires, quel montant de dommages et intérêts l'entreprise mise en accusation devrait-elle être (le cas échéant) condamnée à payer à titre *punitif*, c'est-à-dire pour la sanctionner et pour dissuader d'autres actions similaires à l'avenir, de sa part ou de la part d'autres entreprises ? (Notez SVP votre réponse dans l'espace laissé en blanc ci-dessous.)

$ _____

L'exemple de Joan Glover est une version légèrement abrégée d'un cas utilisé dans une étude réalisée en 1998 par deux d'entre nous (Kahneman et Sunstein, avec notre ami David Schkade[1]). Nous y avons déjà fait allusion dans le chapitre 8, et nous y revenons en détail ici, parce que nous la considérons comme un bon exemple d'audit de bruit, qui illustre de nombreux thèmes de ce livre.

Le sujet central de ce chapitre est le rôle important de l'*échelle de réponse* comme source omniprésente de bruit. Deux personnes peuvent différer dans leurs jugements du même cas, non pas en raison d'un désaccord sur le fond, mais parce qu'elles n'utilisent pas de la même façon l'échelle proposée pour formuler leurs réponses. Quand vous évaluez la performance d'un employé au travail sur une échelle de 0 à 6, et que vous lui donnez un 4, vous pensez peut-être que c'est une assez bonne note. Mais un autre évaluateur pourrait estimer que la même performance vaut un 3, parce que, à ses yeux, 3 est aussi une assez bonne note.

L'ambiguïté des termes utilisés pour graduer les échelles est un problème général. De nombreux chercheurs ont étudié les problèmes de communication qui surviennent quand on utilise des expressions comme « coupable au-delà d'un doute raisonnable[2] », « preuves claires et convaincantes », « performance exceptionnelle » ou « événement improbable[3] ». Les jugements exprimés de cette manière sont inévitablement sujets au bruit parce qu'ils peuvent être interprétés de

manière très variable par les différentes personnes qui les emploient, comme par les différentes personnes qui les entendent.

L'étude qui utilisait le cas Joan Glover visait à observer les effets d'une échelle de réponse ambiguë dans un contexte où cette ambiguïté était lourde de conséquences. Le sujet de l'étude était le bruit dans les dommages et intérêts octroyés par un jury à titre punitif. Comme vous l'avez peut-être deviné à la lecture de la troisième question, la législation aux États-Unis (et dans quelques autres pays) permet aux jurys des procès au civil d'infliger des dommages et intérêts à titre punitif, lorsque les agissements d'un prévenu ont été particulièrement révoltants. Ces dommages et intérêts punitifs s'ajoutent aux dommages et intérêts compensatoires, qui sont destinés à réparer le préjudice subi par la victime. Quand, comme dans l'exemple de Joan Glover, des plaignants assignent en justice une entreprise en raison du préjudice que leur a causé un de ses produits, et quand ils ont gain de cause, ils se voient octroyer une somme d'argent à titre compensatoire, par exemple pour payer leurs frais médicaux et les salaires qu'ils ont perdus. Mais ils peuvent aussi obtenir une indemnité punitive, destinée à sanctionner l'entreprise fautive et à dissuader d'autres entreprises qui seraient tentées de l'imiter. (Dans un procès au pénal, ce serait la fonction d'une amende.) Le cas de General Assistance, dont le comportement était manifestement très répréhensible, est le type même de l'affaire où un jury pourrait raisonnablement attribuer des dommages et intérêts punitifs.

L'un des problèmes que soulèvent les dommages et intérêts punitifs est leur imprévisibilité : le système est sujet à une quantité considérable de bruit. Les demandes de dommages et intérêts punitifs sont souvent rejetées, et même quand elles sont satisfaites, leurs montants, modestes, n'ajoutent souvent pas grand-chose aux dommages et intérêts compensatoires. Mais il y a des exceptions frappantes, où les sommes astronomiques que les jurys accordent semblent aussi surprenantes qu'arbitraires. Un exemple souvent

cité est celui des 4 millions de dollars qu'un concessionnaire automobile a été condamné à payer parce qu'il avait omis de dire au plaignant que sa BMW neuve avait été repeinte[4].

Dans notre étude, il avait été demandé à huit cent quatre-vingt-dix-neuf participants d'évaluer l'affaire Joan Glover et neuf autres affaires similaires, impliquant toutes un plaignant qui avait subi un préjudice et une entreprise responsable de ce préjudice. Contrairement à vous qui avez répondu à trois questions différentes (le niveau d'indignation, l'intention de sanctionner, les dommages et intérêts en dollars), les participants ne devaient répondre qu'à l'une de ces trois questions, mais ils devaient le faire pour chacun des dix cas. Les participants étaient en outre divisés en groupes plus petits, auxquels étaient soumises des versions légèrement différentes de chaque cas. Les différences portaient sur le préjudice plus ou moins grave subi par le plaignant et sur la taille plus ou moins importante de l'entreprise mise en accusation. Ces variations produisaient un total de vingt-huit scénarios. L'objectif de l'étude était double : tester une théorie sur la psychologie des dommages et intérêts punitifs et mesurer le bruit que contribue à créer l'échelle monétaire (ici, le jugement demandé en dollars).

## *L'hypothèse de l'indignation*

Depuis des siècles, philosophes et juristes débattent de ce qui constitue un « juste » châtiment. L'hypothèse que nous voulions tester dans cette étude est que cette question épineuse aux yeux des intellectuels est plutôt simple pour les gens ordinaires, grâce aux mécanismes de la substitution et de l'équivalence. Les jurés, supposions-nous, simplifient l'exercice en substituant à la question difficile une question plus facile, celle à laquelle vous avez immédiatement répondu quand vous avez lu la description des agissements de General Assistance : « Quel est mon degré de colère, ou

d’indignation ? » Il suffit ensuite de chercher une sanction dont l’intensité soit équivalente à celle de l’indignation ressentie.

Pour tester cette « hypothèse de l’indignation », nous avons demandé à des groupes différents de participants de répondre soit à la question de l’indignation, soit à celle de l’intention de sanctionner. Nous avons ensuite comparé la moyenne des réponses données à ces deux questions pour les vingt-huit scénarios de l’étude. La corrélation entre les réponses moyennes aux deux questions était proche de la perfection : 0,98 (PC = 94 %). Cette corrélation suggère que les personnes à qui l’on pose la deuxième question répondent en fait à la première. Le résultat valide l’hypothèse de l’indignation : l’indignation ressentie est le principal déterminant de l’intention de sanctionner[5].

Le principal déterminant… mais pas le seul. Y a-t-il quelque chose dans l’histoire de Joan qui a attiré davantage votre attention quand vous avez noté votre intention de sanctionner que lorsque vous avez évalué votre niveau d’indignation ? Si oui, nous pensons qu’il s’agit du préjudice qu’elle a subi. On peut s’indigner d’un comportement même si l’on n’en connaît pas les conséquences ; dans notre exemple, le comportement de General Assistance était certainement scandaleux avant même que Joan n’en soit la victime. En revanche, les intuitions sur l’intention de sanctionner comportent un élément de rétorsion, voire de vengeance : le châtiment doit correspondre au préjudice subi, ce qu’exprime grossièrement le principe « œil pour œil, dent pour dent ». Ce besoin de proportionnalité dans le châtiment explique que le meurtre et la tentative de meurtre soient traités différemment par la loi et par les jurys : un meurtrier en puissance qui a eu la chance de rater sa cible sera souvent puni moins sévèrement.

Pour savoir si le préjudice subi modifie effectivement l’intention de sanctionner (mais pas le niveau d’indignation), nous avons soumis à différents groupes de participants des versions « préjudice grave » et « préjudice léger » de l’affaire Joan Glover et de plusieurs autres. La

version « préjudice grave » est celle que vous avez lue. Dans la version « préjudice léger », le préjudice était décrit de la manière suivante : Joan « avait dû passer plusieurs jours à l'hôpital et était désormais profondément traumatisée par les cachets et pilules de toutes sortes. Quand ses parents essayaient de lui faire prendre des médicaments comme des vitamines ou de l'aspirine, elle pleurait sans cesse et disait qu'elle avait peur ». Cette version décrit certes une expérience traumatisante, mais le préjudice est bien inférieur à celui de la première version, avec sa longue hospitalisation et ses sérieuses séquelles. La comparaison entre les scénarios permet donc de tester l'effet propre du préjudice subi.

Les résultats furent conformes aux hypothèses que nous avons décrites. À la question de l'indignation, les réponses moyennes étaient presque identiques dans les versions « préjudice grave » et « préjudice léger » (4,24 et 4,19, respectivement). C'est bien le comportement de General Assistance qui suscite l'indignation, pas ses conséquences. En revanche, à la question de l'intention de sanctionner, on constatait une différence statistiquement significative entre les versions grave et légère (respectivement 4,93 et 4,65). Et les dommages et intérêts accordés étaient nettement différents : leur valeur médiane était de 2 millions de dollars pour la version grave et 1 million pour la version légère. Les autres cas soumis aux participants donnèrent des résultats similaires.

Ces résultats mettent en évidence un aspect essentiel du processus du jugement : le contenu du jugement demandé influence la *pondération* des différentes informations fournies. Les sujets de cette étude n'avaient pas conscience qu'ils prenaient position sur la question philosophique de savoir si la sanction doit être proportionnée au préjudice. Ils n'étaient pas même conscients d'attribuer des pondérations différentes aux divers aspects de l'affaire. Mais pour évaluer leur indignation, ils ne tenaient pratiquement aucun compte du préjudice subi par la victime, tandis que pour fixer la sanction, ils donnaient au même préjudice un poids considérable. Rappelons

que les participants ne connaissaient qu'une seule version de l'affaire : le choix d'une sanction plus sévère quand le préjudice était grave ne procédait donc pas d'une comparaison explicite. Il résultait des équivalences établies entre les données du problème d'une part et le jugement demandé d'autre part. Les réponses des participants reposaient davantage sur leur pensée rapide que sur leur pensée lente.

## *Bruyantes échelles*

Le second objectif de l'étude était de démontrer que les dommages et intérêts punitifs sont sujets au bruit. Notre hypothèse était que les jurés sont généralement d'accord sur la sévérité avec laquelle ils souhaitent punir le prévenu, mais qu'ils diffèrent grandement sur la manière de traduire en dollars cette intention de sanctionner.

Le design de l'étude permet de comparer le niveau de bruit dans les jugements des mêmes affaires sur les trois échelles : l'indignation, l'intention de sanctionner et les dommages et intérêts. Pour mesurer le bruit, nous appliquons la même méthode que lorsque nous analysions les résultats de l'audit de bruit des juges fédéraux au chapitre 6. Nous supposons donc, comme nous l'avions fait alors, que, sur chaque cas, la moyenne des jugements de tous les participants peut être considérée comme une « juste peine », exempte de biais. (Rappelons que cette hypothèse peut tout à fait être fausse, mais ne change rien à l'analyse du bruit.) Une fois cette hypothèse posée, tous les jurés qui répondent sur l'une des trois échelles devraient, dans un monde idéal, être d'accord dans leurs jugements de chaque cas. Tout écart par rapport au jugement moyen est défini comme une erreur, et la variabilité de ces erreurs constitue le bruit systémique.

Comme au chapitre 6, ce bruit systémique peut être décomposé en bruit de niveau et en bruit de pattern. Il existe des différences entre les jurés dans

leur niveau moyen de sévérité : certains sont plus enclins que d'autres, sur l'ensemble des cas qui leur sont soumis, à exprimer une vive indignation, à vouloir punir sévèrement les prévenus, ou à fixer des montants importants de dommages et intérêts. Ces différences constituent le bruit de niveau. Le bruit de pattern, quant à lui, survient à cause des réactions variables de chaque juré à chaque cas, indépendamment de la sévérité moyenne du juré en question. Nous pouvons donc décomposer la variance totale des jugements en trois éléments :

**Variance des jugements = variance des justes peines + (bruit de niveau)$^2$ + (bruit de pattern)$^2$**

Cette décomposition est applicable aux trois échelles utilisées : le degré d'indignation, l'intention de sanctionner, et les dommages et intérêts accordés. Mais le résultat est très différent d'une échelle à l'autre, comme le montre la figure 13[6]. L'échelle la moins sujette au bruit est celle de l'intention de sanctionner, où le bruit systémique (la somme des deux composantes grisées sur la figure) représente 51 % de la variance. Ce n'est pas négligeable : il y a à peu près autant de bruit que de justice. Mais il y a nettement plus de bruit sur l'échelle de l'indignation : 71 % de la variance totale. Et l'échelle des dommages et intérêts en dollars est de loin la pire : 94 % de la variance des jugements n'est que du bruit !

Si ces différences sont frappantes, c'est parce que les trois échelles mesurent, en principe, les réponses à des questions presque identiques. Nous avons vu, en présentant l'hypothèse de l'indignation, qu'il existe une corrélation quasi parfaite entre les valeurs moyennes (présumées « justes ») de l'indignation et de l'intention de sanctionner. Quant aux évaluations de l'intention de sanctionner et des dommages et intérêts, elles répondent précisément à la même question (celle de la sévérité avec laquelle General Assistance devrait être punie), et seules les unités dans lesquelles la réponse

est formulée sont différentes. Dès lors, comment pouvons-nous expliquer les énormes différences qu'illustre la figure 13 ?

Un premier constat qu'il est facile de faire est que l'échelle de mesure de l'indignation n'est pas très précise. On peut sans doute qualifier bien des comportements de « totalement acceptables », mais à l'autre extrémité de l'échelle, la limite au degré de colère ou d'indignation que suscitent les agissements de General Assistance (ou d'autres prévenus) est plutôt vague, à supposer qu'elle existe. Que signifie exactement un comportement « absolument scandaleux » ? Ce manque de clarté à l'extrémité supérieure de l'échelle produit inévitablement du bruit.

L'échelle de l'intention de sanctionner est un peu plus précise. Demander « une sanction extrêmement sévère » est moins vague que juger un comportement « absolument scandaleux », parce qu'il y a des limites à ce que peut être une sanction extrêmement sévère. Vous pouvez souhaiter qu'un coupable soit puni de la peine maximum prévue par la loi, mais pas plus : pour scandalisé que vous soyez, vous ne recommanderez pas que le P-DG de General Assistance et toute son équipe dirigeante soient mis à mort (du moins, nous l'espérons). Autrement dit, quand on évalue l'intention de sanctionner, l'échelle est moins ambiguë que quand on exprime son indignation, parce que la borne supérieure de l'échelle est plus claire. Logiquement, cette échelle est aussi moins sujette au bruit.

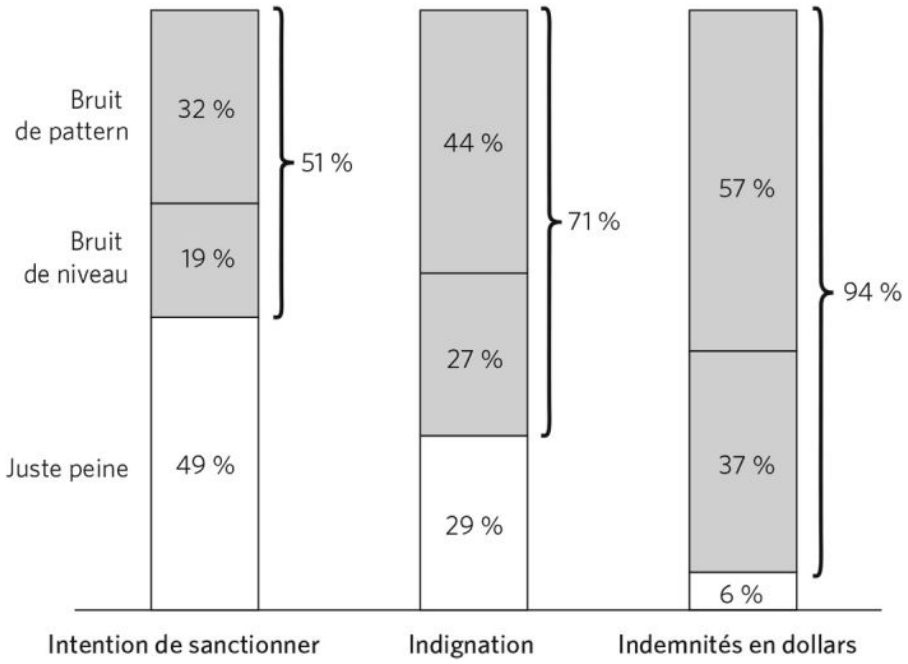



**Figure 13.** Décomposition de la variance des jugements.

L'indignation et l'intention de sanctionner utilisent toutes deux des échelles d'évaluation dont les degrés sont verbaux. L'échelle des dommages en dollars appartient à une famille différente, qui est bien plus problématique.

## *Des dollars et des ancres*

Le titre de notre article de 1998 résume sa principale conclusion : « Indignation partagée et indemnités erratiques ». Nos jurés expérimentaux étaient assez d'accord entre eux dans leur intention de sanctionner ; et leurs évaluations s'expliquaient principalement par leur degré d'indignation, lui aussi partagé. Mais le jugement qui se rapproche le plus de celui qu'on leur aurait demandé dans un véritable tribunal est celui qui consiste à fixer le

montant en dollars des dommages et intérêts. Et dans ce jugement-là, le niveau de bruit était inacceptable.

La raison n'a rien de mystérieux. Si vous vous êtes prêté au jeu en écrivant le montant de dommages et intérêts en dollars qui vous semblait approprié dans l'affaire Joan Glover, vous avez certainement éprouvé le sentiment de faire un choix assez arbitraire. Ce sentiment d'arbitraire recèle une information importante : il implique que les décisions tout aussi arbitraires d'autres personnes seront très différentes, et qu'on peut donc s'attendre à beaucoup de bruit. Comme nous allons le voir, ce niveau élevé de bruit est une caractéristique de la famille d'échelles à laquelle appartient celle des dommages et intérêts en dollars.

On doit à Stanley Smith Stevens, un grand psychologue qui enseigna principalement à Harvard, une découverte étonnante : les gens partagent des intuitions fortes sur les *rapports d'intensités* d'un grand nombre d'expériences et d'attitudes[7]. Ils peuvent ajuster une lampe de sorte qu'elle paraisse « deux fois plus lumineuse » qu'une autre, et ils sont d'accord pour dire que la charge émotionnelle d'une peine de dix mois de prison n'est pas dix fois plus lourde que celle d'une peine d'un mois. Stevens baptisa les échelles qui permettent d'exprimer ce genre d'intuitions des *échelles de rapport*.

Nos intuitions sur l'argent sont exprimées en rapports, ce qui nous permet de comprendre sans difficulté des phrases comme « Sara a eu une augmentation de 60 % » ou « Notre riche voisin a perdu la moitié de sa fortune en une nuit ». L'échelle des dommages punitifs en dollars est donc bien une échelle de rapport, utilisée pour mesurer l'intention de sanctionner. Comme toute échelle de rapport, elle comporte un point zéro significatif (zéro dollar), mais elle n'a pas de limite supérieure.

Stevens a découvert qu'il était possible de calibrer une échelle de rapport en y introduisant une valeur intermédiaire qui permet d'ancrer les jugements. Dans son laboratoire, il exposait ses sujets à une lumière d'une

certaine intensité. Il leur donnait alors pour instruction de « considérer cette luminosité comme une valeur de 10 (ou 50, ou 200) et de chiffrer les autres luminosités en fonction » de ce premier chiffre. Les valeurs que les participants attribuaient ensuite à des lumières de différentes intensités étaient bien proportionnelles à l'ancre arbitraire qu'on leur avait demandé d'adopter. Un sujet ancré sur le chiffre 200 produisait des chiffres vingt fois plus élevés qu'un autre ancré sur le chiffre 10 ; l'écart-type des jugements d'un sujet était, lui aussi, proportionnel à l'ancre fournie.

Au chapitre 13, nous avons décrit un exemple d'ancrage amusant : le prix que vous seriez prêt à payer pour un objet peut être fortement influencé par un nombre arbitraire, composé des deux derniers chiffres de votre numéro de Sécurité sociale. Un résultat plus frappant encore de la même étude est que cette ancre initiale affecte aussi le prix qu'on est prêt à payer pour toute une série d'autres objets. Une fois que les participants avaient consenti à payer pour une souris d'ordinateur une somme relativement importante (par rapport aux autres participants), ils consentaient à payer une somme proportionnellement aussi importante pour un clavier. Nous sommes généralement beaucoup plus sensibles à la valeur *relative* de biens comparables qu'à leur valeur absolue. Cette persistance de l'effet produit par l'ancre initiale produit ce que les auteurs de l'étude appellent de la « cohérence dans l'arbitraire[8] ».

Pour prendre conscience de l'effet qu'une ancre arbitraire pourrait produire dans l'affaire Joan Glover, imaginons que le texte au début de ce chapitre ait contenu les informations suivantes :

> *Dans une affaire similaire impliquant une autre firme pharmaceutique, la petite fille qui en était victime souffrait de traumatismes psychologiques légers [dont la description est celle de la version « préjudice léger » que vous avez lue auparavant]. Les dommages et intérêts punitifs furent fixés à 1,5 million de dollars.*

Le problème du choix d'une sanction pour General Assistance est tout d'un coup devenu beaucoup plus facile ! Un montant vous est peut-être même déjà venu à l'esprit. Le contraste entre le préjudice grave infligé à Joan et le préjudice léger subi par l'autre petite fille est un rapport. Ce rapport devient intuitivement un multiplicateur que vous pouvez appliquer à des dommages et intérêts en dollars. Qui plus est, l'ancre initiale proposée ici (la somme de 1,5 million de dollars) sera suffisante pour ancrer toute l'échelle des sanctions en dollars dans d'autres cas. Pour fixer des dommages et intérêts dans des affaires plus graves que les deux considérées jusqu'ici (ou moins graves qu'elles), il vous suffira de vous demander *dans quel rapport* elles sont plus graves (ou moins graves) que les précédentes.

Si nous avons besoin d'une ancre pour produire des jugements sur une échelle de rapport, que se passe-t-il quand nous n'en avons pas ? Stevens a donné la réponse. En l'absence d'instructions données par l'expérimentateur, les sujets qui utilisent une échelle pour la première fois sont contraints de faire un choix arbitraire. À partir de là, il leur est possible de produire des jugements cohérents, en utilisant leur première réponse comme ancre.

Vous le voyez maintenant : l'exercice que vous avez fait au début de ce chapitre, quand vous deviez fixer un montant de dommages et intérêts dans l'affaire Joan Glover, était un exemple de ce type de problème. Nous vous y demandions d'utiliser une échelle de rapport sans vous fournir d'ancre. Comme les sujets dépourvus d'ancre dans le laboratoire de Stevens, vous avez dû fixer arbitrairement le montant de la sanction.

Les participants à notre étude ont rencontré le même problème. Eux aussi ont été contraints de proposer un chiffre arbitraire sur la première affaire qui leur était soumise. Contrairement à vous, cependant, ils ne se sont pas arrêtés là, puisqu'on leur demandait par la suite de fixer des dommages punitifs dans neuf autres affaires. Les neuf jugements suivants n'étaient pas arbitraires : chaque participant pouvait les fixer de manière à

rester cohérent avec son jugement initial, s'assurant ainsi que les jugements soient cohérents les uns avec les autres. Chacun pouvait rester cohérent dans l'arbitraire.

On peut déduire des principes découverts par Stevens une hypothèse quant au résultat de ce processus : l'ancre fixée dans la première affaire doit influencer fortement les valeurs absolues des jugements en dollars dans les neuf autres affaires, mais elle ne devrait avoir absolument aucun effet sur les positions *relatives* des dix affaires. Par exemple, si un juré accorde, dans la première affaire, une somme très importante, toutes les autres sommes accordées devraient s'en trouver accrues en proportion, mais les rapports entre ces sommes ne devraient pas être affectés. Ce raisonnement conduit à une conclusion surprenante : en dépit du bruit considérable qu'ils contiennent, les jugements en dollars doivent refléter les véritables intentions de sanctionner. Pour révéler ces intentions et retrouver le « signal » noyé dans le bruit des jugements en dollars, il doit suffire de remplacer ces valeurs absolues par des valeurs relatives, qui comparent entre eux les jugements d'un même juré sur différentes affaires.

Pour tester cette hypothèse, nous avons refait l'analyse du bruit en remplaçant chaque montant de dommages et intérêts en dollars par son rang dans les dix jugements d'un individu : les dommages et intérêts les plus élevés accordés par un juré donné étaient notés 1, les suivants 2, et ainsi de suite. Cette transformation des montants en rangs dans un classement supprime tout bruit de niveau entre les jurés, puisque les rangs notés de 1 à 10 sont les mêmes pour tous (sauf dans les quelques cas où le classement comprend des *ex aequo*[9]).

Les résultats furent spectaculaires : la part du bruit dans la variance des jugements chuta de 94 à 49 % (figure 14). La transformation des montants monétaires en rangs dans le classement de ces montants montra que les jurés étaient en réalité largement d'accord sur la sanction appropriée dans les différentes affaires[10]. Dans leurs classements, il y avait même

légèrement *moins* de bruit que dans les intentions de sanctionner mesurées séparément.

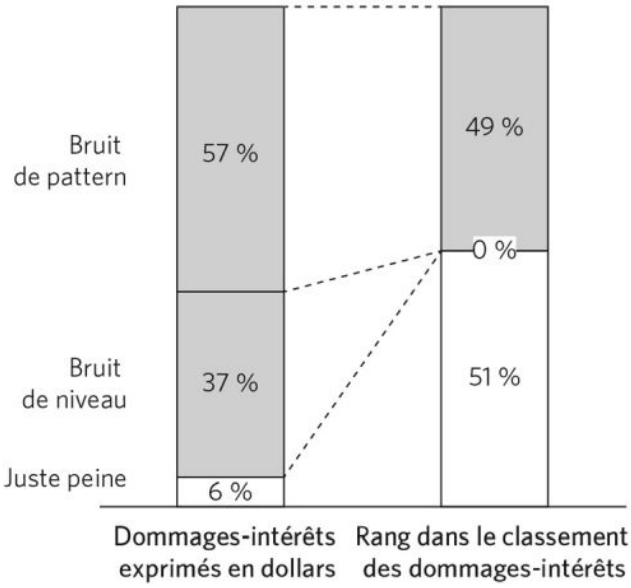



**Figure 14.** Bruit dans les montants et bruit dans leur classement.

## *Une conclusion préoccupante*

Ces résultats sont cohérents avec la théorie que nous avons proposée : pour toutes les affaires, les dommages et intérêts en dollars étaient ancrés sur le chiffre arbitrairement choisi par chaque juré dans la première affaire examinée. Le classement relatif des cas reflète assez fidèlement les attitudes des jurés, et n’est donc pas très affecté par le bruit. Les montants absolus des dommages et intérêts en dollars, en revanche, n’ont à peu près aucun sens, car ils dépendent du montant arbitraire choisi dans la première affaire.

Paradoxalement, dans les procès réels, la première affaire que les jurés ont à juger est aussi la dernière. Aux États-Unis, on attend des jurys dans les procès au civil qu’ils fixent un montant de dommages et intérêts en

dollars pour l’affaire qui leur est soumise, sans pouvoir bénéficier du moindre ancrage pour s’orienter. La loi interdit expressément toute communication au jury sur les montants des dommages et intérêts punitifs accordés dans d’autres affaires. L’hypothèse implicite sur laquelle repose cette loi est que le sens de la justice des jurés leur permet de passer directement de leur point de vue sur le délit à la fixation d’une juste peine. Cette hypothèse est un non-sens psychologique : elle suppose une capacité dont les humains sont dépourvus. Les institutions judiciaires seraient bien inspirées de tenir compte des limites des individus qui administrent la justice.

---

L’exemple des dommages punitifs aux États-Unis est extrême : il est rare qu’on impose à des professionnels des échelles aussi ambiguës (et plus rare encore qu’on interdise formellement de dissiper cette ambiguïté !). Néanmoins, les échelles ambiguës sont courantes.

L’étude sur les dommages et intérêts punitifs nous enseigne deux leçons générales sur les échelles, qu’on peut appliquer dans l’entreprise, dans l’enseignement, dans le sport, dans l’administration, ou n’importe où ailleurs. Premièrement, le choix de l’échelle qui sert à formuler les jugements est essentiel : plus les échelles sont ambiguës, plus elles sont sources de bruit. Deuxièmement, quand il est possible de remplacer des jugements absolus par des jugements relatifs ou ordinaux, il est très probable que cela contribue à réduire le bruit.

## À propos des échelles

« Il y a beaucoup de bruit dans nos jugements. Est-il possible que nous ayons une interprétation différente des niveaux utilisés sur l'échelle de réponse ? »
« Puisque nous devons juger d'une série d'affaires différentes, pouvons-nous nous mettre d'accord sur une affaire de référence, afin que ce jugement fasse office d'ancre quand nous examinerons les affaires suivantes ? »
« Pour réduire le bruit, aurions-nous intérêt à remplacer nos jugements en valeur absolue par un classement ? »

CHAPITRE 16

# Patterns

Vous souvenez-vous de Julie, l'enfant dont, au chapitre 14, vous avez essayé de deviner la note moyenne à l'université (GPA) ? En voici une description plus complète :

> *Julie était fille unique. Son père était un avocat réputé, sa mère était architecte. Quand Julie avait 3 ans, son père contracta une maladie auto-immune qui l'obligea à travailler à domicile. Il passait beaucoup de temps avec Julie, et c'est lui qui, patiemment, lui apprit à lire. Julie lisait couramment à l'âge de 4 ans. Son père essaya aussi de lui apprendre l'arithmétique, mais avec moins de succès. À l'école élémentaire, Julie avait de bons résultats scolaires, mais elle avait du mal à s'intégrer et n'était pas très appréciée. Assez solitaire, elle se prit de passion pour l'observation des oiseaux, le hobby de son oncle préféré.*
>
> *Ses parents divorcèrent quand elle avait 11 ans, et Julie le vécut très difficilement. À l'école, ses notes dégringolèrent, et ses fréquents éclats de colère lui valurent des sanctions disciplinaires. Plus tard, au lycée, elle réussit très bien dans certaines matières, dont la biologie et l'écriture créative. Elle surprit aussi tout le monde en obtenant*

*d'excellents résultats en physique-chimie. Mais elle négligea la plupart des autres matières, et sortit du lycée avec une moyenne générale de « B » [équivalente à un GPA de 3,0].*

*N'étant pas admise dans les écoles prestigieuses auxquelles elle aspirait, Julie finit par entrer dans une bonne université d'État, où elle se spécialisa d'abord dans les études environnementales. Ses deux premières années furent assombries par des relations amoureuses multiples et compliquées, ainsi que par une consommation régulière de marijuana. Mais à partir de la fin de sa deuxième année, elle conçut un fort désir d'intégrer une école de médecine et, pendant les deux années suivantes, travailla beaucoup plus sérieusement.*

À votre avis, quelle note moyenne générale Julie a-t-elle obtenue à son diplôme au terme de ses quatre années d'université ?

## *Problèmes faciles et problèmes difficiles*

À l'évidence, ce problème (appelons-le Julie 2.0) est devenu beaucoup plus difficile. Tout ce que vous saviez de Julie 1.0 était qu'elle avait su lire à l'âge de 4 ans. Avec un seul indice, il était facile d'établir une équivalence, et une estimation intuitive de sa note moyenne générale vous venait rapidement à l'esprit.

Une équivalence serait encore possible si vous aviez plusieurs indices allant tous globalement dans la même direction. Par exemple, dans la description de Bill, le comptable joueur de jazz, toutes les informations (« sans imagination », « fort en mathématiques », « faible dans les matières littéraires ») brossaient un tableau cohérent, voire stéréotypé. De même, si la plupart des événements dans la vie de Julie 2.0 correspondaient au cliché d'une enfant précoce et brillante (avec éventuellement une ou deux données

simplement « moyennes »), vous ne trouveriez pas l'exercice si difficile. Quand les éléments disponibles composent un tableau cohérent, la pensée rapide de notre Système 1 n'a aucun mal à lui donner un sens. Un tel tableau caractérise un problème de jugement simple, facilement résolu, et dont la plupart des gens s'accordent sur la solution.

Mais Julie 2.0 n'est pas un jugement simple. Ce qui le rend difficile, c'est la présence de nombreux indices contradictoires. L'histoire qui nous est racontée comprend des signes d'aptitude et de motivation, mais suggère aussi des faiblesses de caractère, et mentionne des résultats médiocres. L'histoire, d'ailleurs, n'en est pas une : elle semble aller dans tous les sens. On n'en saisit pas la conclusion, parce que ses éléments ne permettent pas de former une interprétation cohérente. Notons au passage que cette incohérence n'a rien d'invraisemblable. La vie est souvent plus complexe que les histoires que nous aimons raconter.

La présence d'indications multiples et contradictoires est source d'ambiguïté, et cette ambiguïté caractérise les problèmes de jugement difficiles. L'ambiguïté explique aussi que les problèmes difficiles soient plus sujets au bruit que les problèmes simples ; car s'il existe plus d'une manière de considérer quelque chose, on peut être certain que les gens vont varier dans leur manière de le considérer. Dans notre exemple, différents lecteurs choisiront des éléments différents pour former le noyau de leur récit, et en ignoreront d'autres. Il y aura donc un grand nombre d'interprétations possibles. La difficulté que vous avez éprouvée à trouver une cohérence dans l'histoire de Julie 2.0 est précisément ce qui garantit que d'autres lecteurs, confrontés à cette même difficulté, la résoudront en trouvant une cohérence différente et arriveront à une autre conclusion que la vôtre. C'est cette variabilité qui fait naître le bruit de pattern.

L'existence de jugements difficiles soulève une question importante : celle du degré de confiance que nous avons dans nos jugements. Quelle est la différence entre un jugement sur lequel vous hésitez et un jugement dans

lequel vous avez une grande confiance ? Pour que vous soyez confiant, deux conditions doivent être réunies. D'abord, l'histoire à laquelle vous croyez doit être globalement cohérente. Cette cohérence globale exige que tous les détails de l'interprétation retenue correspondent à l'histoire et se renforcent les uns les autres. À défaut, on peut obtenir cette cohérence moins élégamment, en négligeant ou en trouvant des raisons d'ignorer les détails qui ne « collent » pas avec l'histoire.

Ensuite, il ne doit pas y avoir d'autre interprétation possible des faits qui soit aussi séduisante que la vôtre. Le véritable expert qui a « résolu » un problème de jugement est non seulement capable de justifier son interprétation des faits, mais aussi d'expliquer de manière persuasive pourquoi les autres histoires qu'on pourrait proposer sont fausses. Là encore, on peut prendre confiance dans son propre jugement en se masquant une partie de la difficulté : c'est ce que l'on fait si l'on n'envisage pas les interprétations alternatives, ou si l'on les exclut activement.

La principale conséquence de cette conception de la confiance est que la confiance subjective dans un jugement n'en garantit nullement l'exactitude. En outre, la suppression des interprétations alternatives – un processus perceptif aujourd'hui bien établi [1] – contribue à ce que nous avons appelé l'illusion d'accord (voir chapitre 2). En effet, quelqu'un qui n'imagine pas d'alternatives possibles à sa conclusion supposera naturellement que tous les autres observateurs du même problème doivent arriver eux aussi à la même conclusion.

Bien sûr, rares sont les individus qui ne connaissent pas le doute. Nous savons tous ce qu'est l'incertitude (il suffit pour s'en souvenir de relire le cas de Julie 2.0). Nous n'avons pas tous, à tout instant, une confiance totale dans nos jugements. Mais généralement, nous sommes plus confiants que nous ne devrions l'être [2].

## *Le bruit de pattern, occasionnel ou stable*

Nous avons défini l'erreur de pattern comme l'erreur de jugement d'un individu sur une affaire que la disposition générale de cet individu (son erreur de niveau) ne suffit pas à expliquer. C'est par exemple le cas d'un juge généralement clément, mais qui se montre d'une sévérité inhabituelle quand il a affaire à un multirécidiviste, ou quand le délit est un grand excès de vitesse. C'est aussi ce qui arrive à l'investisseur normalement prudent qui se laisse séduire, contre toute attente, par le *business plan* d'une start-up prometteuse. Bien sûr, les erreurs de pattern peuvent aussi être moins visibles, quand elles ne vont pas à l'encontre de la tendance générale, mais, au contraire, l'accentuent : on observera une erreur de pattern modérée (et discrète) quand un juge réputé indulgent se montre encore plus clément que d'habitude face à une jeune femme.

Certains des facteurs qui causent les erreurs de pattern sont momentanés : ils affectent une occasion de jugement, mais n'affecteraient pas forcément le même individu s'il examinait le même problème à un autre instant. Ce sont toutes les sources de ce que nous avons appelé le bruit occasionnel : par exemple, la bonne humeur d'un individu à l'instant du jugement, une contrariété récente qu'il a encore à l'esprit mais qu'il aura oubliée demain, ou son état de fatigue.

Mais d'autres facteurs d'erreurs de pattern peuvent être considérés comme permanents. Certains employeurs, par exemple, sont particulièrement (et durablement) favorables aux diplômés de telle ou telle université. Et certains médecins, quand ils examinent un patient qui souffre de pneumonie, peuvent avoir une plus grande propension que leurs confrères à recommander une hospitalisation. Ces tendances ne sont pas fonction de l'humeur du moment : on peut s'attendre au contraire à ce qu'elles soient stables dans le temps.

Les effets de ces facteurs stables s'ajoutant à ceux des facteurs momentanés, l'erreur de pattern observée dans un jugement peut être décrite par cette équation :

**Erreur de pattern = erreur de pattern stable + erreur occasionnelle**

Comme dans les précédentes décompositions d'erreur que nous avons proposées, et grâce au fait que l'erreur de pattern stable et l'erreur occasionnelle sont indépendantes et non corrélées, nous pouvons étendre cette équation à l'analyse des variances de ces erreurs, et donc décomposer le bruit de pattern :

$$\textbf{(Bruit de pattern)}^2 = \textbf{(bruit de pattern stable)}^2 + \textbf{(bruit occasionnel)}^2$$

De la même manière, toujours comme pour les autres composantes de l'erreur, cette équation peut être représentée graphiquement comme une somme de carrés dont les côtés sont ceux d'un triangle rectangle (figure 15).

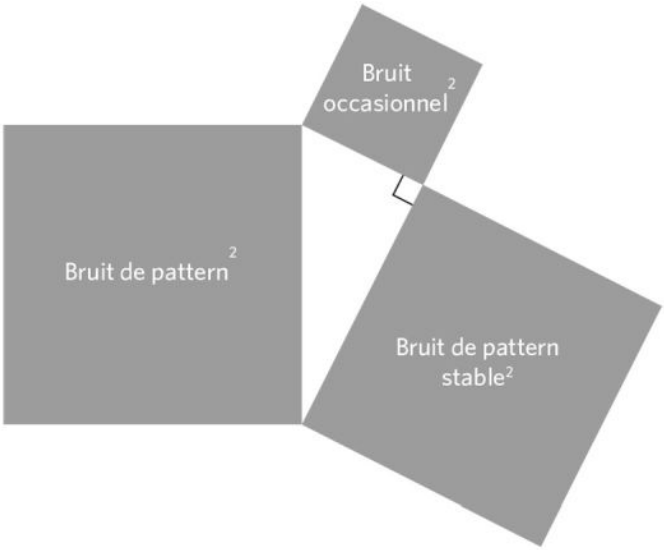



**Figure 15.** Décomposition du bruit de pattern.

## *Aux sources du bruit de pattern stable*

L'idée de *bruit de pattern stable* est moins intuitive que celle de bruit occasionnel ; et on s'en représente moins facilement les sources. Pour l'appréhender, commençons par en donner quelques exemples.

Reprenons par exemple le problème que nous avons présenté au chapitre 9, où des recruteurs s'efforcent de prévoir la performance future de cadres dirigeants, sur la base d'une série de notes. Nous avons décrit la manière dont on peut construire un « modèle du juge », c'est-à-dire une formule qui reproduit assez fidèlement les choix d'un individu donné. Le modèle de chaque recruteur attribue à chaque note (leadership, communication, etc.) un certain coefficient, qui correspond à la pondération de cette dimension dans le jugement du recruteur.

Or ces coefficients varient d'un recruteur à l'autre : ils reflètent le fait que tous les recruteurs n'accordent pas la même importance aux différentes notes. Si vous avez déjà participé à un comité de recrutement, vous avez

peut-être observé vous-même ce type de préférence. Un recruteur peut par exemple considérer que les qualités de leadership sont absolument essentielles, tandis qu'un autre accorde un poids plus important aux compétences en communication. Ces différences sont stables, puisqu'elles sont fonction des préférences des recruteurs, observées et modélisées sur un grand nombre de décisions. Néanmoins, elles produisent de la variabilité dans le classement que les recruteurs font des candidats. Cette *variabilité dans les pondérations accordées aux dimensions du problème* crée donc du bruit de pattern stable.

Le bruit de pattern stable peut aussi provenir de réactions personnelles et extrêmement spécifiques à des cas individuels, plutôt qu'à des dimensions du problème. Songez par exemple à ce qui vous a conduit à prêter davantage attention à certains aspects de l'histoire de Julie qu'à d'autres. Certains détails ont pu résonner avec votre propre expérience : peut-être quelque chose dans l'histoire de Julie vous a-t-il rappelé un proche qui, malgré des facilités intellectuelles évidentes, va d'échec en échec, et dont les faiblesses de caractère vous semblaient manifestes depuis son adolescence. Inversement, l'histoire de Julie a pu évoquer chez vous le souvenir d'une amie qui, malgré une adolescence difficile et une première orientation dans une tout autre voie, a réussi à faire médecine et est aujourd'hui une spécialiste reconnue. Les souvenirs et les associations qu'évoque la description de Julie chez différentes personnes sont imprévisibles et propres à chaque lecteur. Mais ils sont en bonne partie stables : si vous aviez lu l'histoire de Julie la semaine dernière, elle vous aurait rappelé les mêmes personnes et les mêmes histoires, et votre jugement aurait été tout aussi personnel. *Les associations qu'évoque chaque cas* sont propres à chaque juge et à son histoire ; et elles créent elles aussi du bruit de pattern stable.

Une troisième cause possible de bruit de pattern stable est la présence de différences entre les individus dans la justesse de leurs jugements. En

poussant ce raisonnement à l'extrême, imaginons un prévisionniste doté d'extraordinaires pouvoirs de divination, et qui pourrait prédire l'avenir avec une exactitude parfaite. Sa prescience le ferait souvent dévier de la moyenne des prévisions, puisqu'il aurait toujours raison, même quand les autres se trompent. Pourtant, tant que les résultats ne sont pas connus, ces écarts par rapport à la moyenne seraient analysés comme des erreurs de pattern. Ce cas d'école illustre une règle générale : lorsqu'un jugement est invérifiable, une *différence dans les niveaux de compétence* de différents juges crée du bruit de pattern.

Enfin, du bruit de pattern stable peut apparaître lorsque les compétences des juges diffèrent non par leur niveau, mais par leur nature. Des personnes qui ont des spécialités et des aptitudes différentes vont nécessairement prendre en compte des dimensions différentes d'un même problème. Imaginons par exemple que plusieurs experts contribuent au processus de sélection d'un club de sport professionnel. Il y a là des entraîneurs, qui se focalisent sur la technicité des sportifs, des médecins qui évaluent leur susceptibilité aux blessures, et des psychologues qui évaluent leur motivation et leur résilience. La comparaison des évaluations des mêmes joueurs par ces différents spécialistes révélera un bruit de pattern considérable. Une version moins extrême du même problème se présente quand les différents juges en présence ne sont pas des spécialistes différents, mais simplement des généralistes qui se trouvent être plus qualifiés pour évaluer certaines qualités que d'autres. Dans ce type de situation, le bruit de pattern stable n'est pas tant une erreur que la manifestation de *la variété des compétences individuelles des juges*.

Une précision s'impose sur ce dernier cas : la variété des compétences n'est une source de bruit que lorsque les professionnels jugent et décident seuls. Quand, au contraire, on peut constituer des équipes chargées de produire des jugements ensemble, la diversité des compétences devient un atout potentiel : les différents individus peuvent prendre en compte des

aspects différents et complémentaires du problème de jugement auquel ils font face[3]. Nous parlerons de cette possibilité, et de ce qu'il faut faire pour l'exploiter, au chapitre 21.

Dans des chapitres précédents, nous avons parlé des deux « loteries » auxquelles sont confrontés les clients d'une compagnie d'assurances ou les prévenus appelés à comparaître devant un juge. La première loterie choisit dans un groupe de collègues la personne à laquelle un cas donné sera attribué. Elle détermine donc, bien sûr, le niveau moyen des jugements de ce professionnel, c'est-à-dire son erreur de niveau : par exemple, le fait qu'un prévenu va avoir affaire à un juge réputé sévère ou au contraire laxiste. Mais nous voyons maintenant que ce n'est pas tout. En sélectionnant un individu, la première loterie choisit aussi tout le bagage de cet individu : l'ensemble des valeurs, des préférences, des croyances, des souvenirs, des expériences et des associations qui lui sont propres.

Cet ensemble de caractéristiques personnelles compose un pattern unique. Quand vous abordez un nouveau problème de jugement, vous le faites avec vos habitudes de pensée, avec les leçons que vous avez tirées de l'expérience, de votre formation, de l'enseignement de vos mentors. Vous êtes fort des succès qui ont bâti votre confiance, et riche des erreurs que vous avez commises et que vous prenez garde à ne pas répéter. Et quelque part dans votre cerveau se trouvent les règles formelles dont vous vous souvenez, celles que vous avez oubliées, et celles dont vous avez appris à ne plus tenir compte. Personne n'est exactement comme vous dans chacun de ces domaines. Votre pattern personnel est aussi unique que vous. Les erreurs de pattern stables qu'il engendre le sont aussi.

La seconde loterie contribue elle aussi au bruit de pattern, mais pas de la même manière. En déterminant le moment où nous produisons notre jugement, elle détermine l'humeur dans laquelle nous nous trouvons et diverses autres circonstances exogènes. Ces circonstances, nous l'avons vu,

ne devraient pas affecter notre jugement mais l'affectent tout de même : c'est le bruit occasionnel.

Imaginons, par exemple, que juste avant de lire la description de Julie, vous ayez lu dans le journal un article alarmant sur la consommation excessive de drogues à l'université. L'article en question contenait en particulier l'histoire d'un étudiant en droit doué et motivé, mais qui s'était avéré incapable de compenser le déficit de connaissances accumulé pendant ses premières années d'université, à une époque où il était un grand fumeur de marijuana. Fraîchement présente à votre esprit, cette analogie va vous amener à prêter une attention accrue à un élément particulier de l'histoire de Julie (sa consommation de marijuana). Vous allez donner plus de poids à cet élément dans votre évaluation de ses chances. Mais si l'on vous lisait la description de Julie dans quelques semaines, vous ne vous souviendriez sans doute plus de cet inquiétant article (dont, évidemment, vous n'auriez rien su non plus si le problème vous avait été soumis hier). L'effet que produit sur vous l'article de journal est transitoire : c'est du bruit occasionnel.

Cet exemple illustre la différence entre le bruit de pattern stable et son variant instable, le bruit occasionnel ; mais il montre aussi que la frontière entre les deux n'est pas parfaitement étanche. Après tout, quelqu'un d'autre n'aurait pas forcément réagi de la même manière que vous à la lecture du même article ; et vous-même n'en auriez peut-être pas tiré les mêmes conclusions si vous l'aviez lu à un autre moment. La principale différence est de savoir si la sensibilité particulière d'une personne à certains aspects d'un problème de jugement est passagère ou permanente. Quand les éléments déclencheurs du bruit de pattern sont intrinsèquement transitoires (comme la fatigue à la fin d'une longue journée), le bruit est occasionnel. Mais quand ces déclencheurs sont enracinés dans nos valeurs et nos expériences personnelles, nous pouvons nous attendre à ce que le pattern

soit un reflet stable de nos préférences, de nos goûts, de notre histoire – bref, de tout ce qui fait que nous sommes uniques.

## *Personnalité et jugement*

Le bruit de pattern stable est donc le produit d'un ensemble de réactions personnelles et uniques à chaque problème. Il est pourtant stable dans le temps. Ces caractéristiques qui pourraient sembler paradoxales cessent de l'être si l'on se réfère à une analogie : les patterns stables de jugement sont comparables à la personnalité des gens qui nous entourent. Comme sa personnalité, le pattern stable de jugements d'un individu est unique. Il constitue en quelque sorte sa « personnalité de jugement ».

Les questions de jugement auxquelles nous nous intéressons dans ce livre peuvent être replacées dans un contexte plus large. Le fait pour un juge de produire un jugement sur une affaire peut en effet être considéré comme un cas particulier d'une question plus générale : *comment une personne agit-elle dans une situation donnée ?* Cette question est du ressort de la psychologie de la personnalité, qui y a consacré des décennies d'études approfondies.

Il y a longtemps que les psychologues cherchent à comprendre et à mesurer les différences de personnalité entre individus. La liste de ces différences est longue : en 1936, une plongée dans le dictionnaire à la recherche de termes susceptibles de décrire une personne avait permis d'identifier dix-huit mille mots[4]. Aujourd'hui, le modèle dominant de la personnalité, dit modèle en cinq facteurs ou modèle des « Big Five », regroupe les différents traits de personnalité en cinq catégories : ouverture à l'expérience, caractère consciencieux, extraversion, amabilité, et stabilité émotionnelle. Chacune de ces catégories recouvre à son tour un large ensemble de traits distincts.

Un trait de personnalité est considéré comme pertinent dans la mesure où il est un bon prédicteur de comportements réels. Si quelqu'un est qualifié de consciencieux, nous nous attendons à observer chez lui les comportements correspondants : arriver à l'heure, respecter ses engagements, etc. Et si Andrew est noté plus haut que Brad sur une échelle d'agressivité, nous devrions observer qu'Andrew, dans la plupart des situations, se comporte de façon plus agressive que Brad. En réalité, cependant, la validité des traits généraux pour prédire des comportements particuliers est assez limitée. Comme nous l'avons noté en parlant des mesures de corrélation dans les jugements prédictifs, une corrélation de 0,30 (PC = 60 %) serait déjà considérée comme élevée[5].

En particulier, le sens commun suggère que si le comportement peut être mû par la personnalité, il est aussi fortement affecté par les *situations*. Dans certaines situations personne n'est agressif, et dans d'autres tout le monde l'est. Lorsqu'ils consolent un ami en deuil, ni Andrew ni Brad ne sont agressifs ; sur un terrain de rugby, en revanche, tous deux font montre d'une certaine agressivité. En somme – et cela n'a rien d'étonnant – les comportements sont fonction à la fois des personnalités *et* des situations.

Ce qui rend chaque individu unique (et l'étude des personnalités éternellement passionnante) est que cette combinaison de la personnalité et de la situation n'est pas une fonction additive, automatique et prévisible. Certaines situations suscitent évidemment plus d'agressivité *en moyenne*, mais les situations qui provoquent plus ou moins d'agressivité ne sont pas les mêmes pour tout le monde. Ainsi, même si Andrew et Brad font preuve du même niveau d'agressivité moyen, ils ne se comporteront pas pour autant de la même manière dans toutes les situations. Peut-être Andrew est-il agressif avec ses pairs mais docile avec ses supérieurs, tandis que le niveau d'agressivité de Brad est insensible au niveau hiérarchique. Et peut-être Brad devient-il maladivement agressif quand on le critique, tout en restant bizarrement timoré quand il est physiquement menacé[6].

Ces patterns de réactions aux situations sont autant de signatures personnelles et ont tendance à être assez stables dans le temps. Ils constituent une large part de ce que nous considérons comme la *personnalité* de quelqu'un, même s'ils ne se laissent pas réduire à quelques grands traits mesurables. Andrew et Brad pourraient avoir la même note sur un test de personnalité qui mesure l'agressivité, mais ils sont uniques dans leurs patterns respectifs de réaction aux déclencheurs de cette agressivité. Plus généralement, quand on dit que deux personnes ont à peu près le même niveau sur un trait de personnalité donné – qu'elles sont, par exemple, également tenaces ou également généreuses –, on devrait plutôt parler de deux distributions de comportements qui ont la même moyenne. Dans les mêmes situations, ces deux personnes n'ont certainement pas exactement les mêmes comportements. Chacune a un pattern de réactions aux situations qui lui est propre.

Il existe un parallèle frappant entre cette analyse de la personnalité et le modèle du jugement que nous proposons ici. Les différences de niveau moyen entre juges (sources du bruit de niveau) sont analogues aux différences de traits de personnalité : ce sont des moyennes. Quant aux problèmes de jugement, ils sont analogues aux situations dans lesquelles on observe différents comportements. Ainsi, le jugement produit par un individu donné sur un problème particulier dépend à la fois du problème et du niveau moyen des jugements de cet individu ; et pourtant, il n'est que modérément prévisible à partir de ces deux données. C'est la même imprévisibilité qu'on observe (en s'en étonnant sans doute moins) dans le domaine de la personnalité : si vous connaissez le niveau moyen des traits de personnalité d'un individu donné et la réaction moyenne d'autres personnes dans une situation donnée, vous ne serez pas pour autant capable de prévoir avec une grande précision le comportement de cet individu dans cette situation.

Une autre manière de se représenter le bruit de pattern stable est de classer les individus en fonction du niveau de leurs jugements. Si l'on fait cet exercice séparément pour chaque cas, le classement variera beaucoup d'un cas à un autre, parce que les individus diffèrent dans leurs réactions propres à chacun des cas examinés. La « signature » d'un juge est un pattern unique de sensibilité aux composantes des différents cas, qui crée un pattern unique de jugements de ces cas. Ce pattern est aussi unique que sa personnalité. En somme, si nos jugements diffèrent les uns des autres avec une certaine stabilité, c'est parce que nous sommes différents les uns des autres par nos personnalités, nos histoires, et nos préférences, qui ont, elles aussi, une certaine stabilité.

---

Si le bruit occasionnel est le produit des circonstances, le bruit de pattern stable constitue en quelque sorte la signature d'un juge, une marque de sa personnalité. Cette analogie appelle un commentaire important. En général, si l'on évoque la variété infinie des caractères humains, c'est pour s'en émerveiller. On célèbre à juste titre la diversité des profils et des expériences ; et l'on se réjouit volontiers d'être entouré de personnalités diverses.

Pourtant, quand on parle de jugements professionnels, la variabilité indésirable est toujours problématique. L'analogie avec la personnalité nous rappelle que le bruit de pattern dans le jugement est en grande partie stable et non aléatoire (même si nous avons peu d'espoir de l'expliquer). Mais quand on s'attend à des jugements convergents, la variété des patterns de jugement ne peut pas être un atout. Si des jugements sont différents, parce qu'ils reflètent la personnalité et les préférences singulières de ceux qui les formulent, ils ne peuvent pas tous être justes. Le bruit, même quand il provient de ce qui fait de nous des individus uniques, reste toujours de l'erreur.

## À propos du bruit de pattern

« Vous semblez confiant dans votre conclusion, mais ce n'est pas un problème facile : il y a des indices qui vont dans différentes directions. Auriez-vous négligé des interprétations alternatives des éléments dont vous disposez ? »
« Vous et moi avons interviewé la même candidate. Nous avons habituellement un même niveau d'exigence. Et pourtant nos jugements sont complètement différents. D'où vient ce bruit de pattern ? »
« Le caractère unique de chaque individu, la personnalité, l'originalité, rendent les gens capables d'innovation et de créativité, et tout simplement intéressants… Mais quand on est face à un problème de jugement, l'unicité et l'originalité cessent d'être des atouts. »

CHAPITRE 17

# Les composantes du bruit

Nous espérons que vous êtes désormais d'accord pour dire que « dès lors qu'il y a jugement, il y a bruit ». Peut-être même avez-vous cessé de penser qu'il y en a « plus qu'on ne l'imagine ». Ce mantra sur le bruit nous guidait quand nous avons commencé à écrire ce livre, mais notre réflexion sur le sujet a évolué au fil des années que nous lui avons consacrées. Dans ce chapitre, nous passons en revue les principales leçons que nous avons apprises sur les composantes du bruit, sur leur importance respective dans le panorama général du bruit, et sur la place du bruit dans l'étude du jugement.

## *Les composantes de l'erreur*

La figure 16 propose une représentation graphique combinée des trois équations que nous avons introduites dans les chapitres 5, 6 et 16. Elle reprend nos trois décompositions successives de l'erreur :

- la décomposition de l'erreur globale en biais et en bruit systémique ;

- la décomposition du bruit systémique en bruit de niveau et en bruit de pattern ;
- la décomposition du bruit de pattern en bruit de pattern stable et en bruit occasionnel.

Vous pouvez maintenant voir comment l'erreur globale (erreur quadratique moyenne, ou MSE) se compose des carrés du biais et des trois composantes du bruit que nous avons présentées [1].

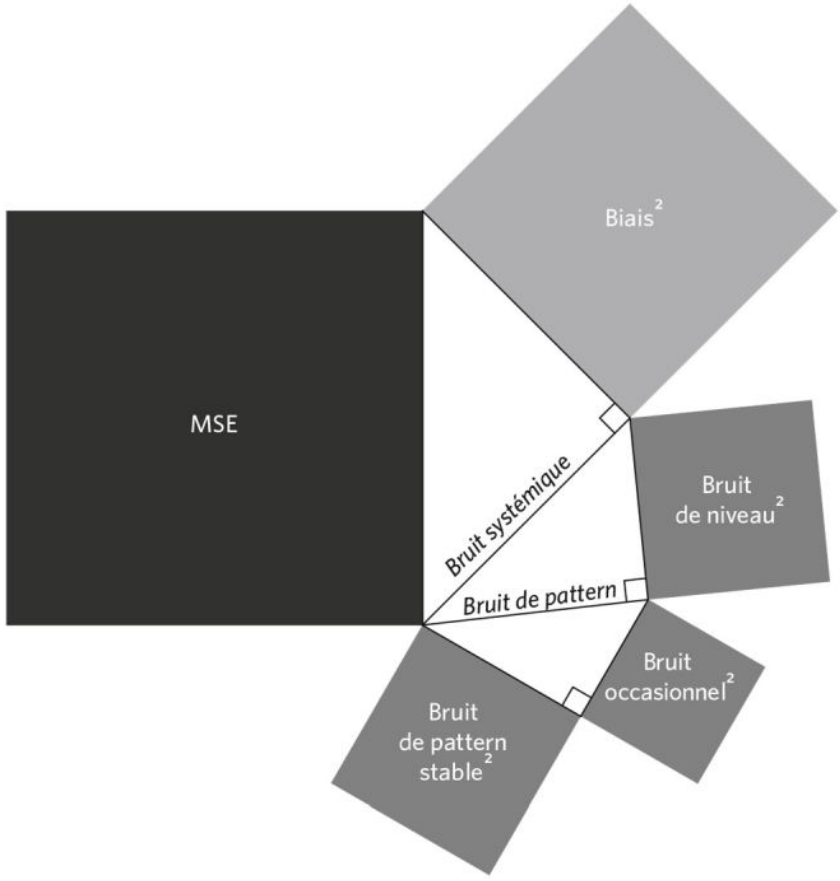



**Figure 16.** L'erreur, le biais et les composantes du bruit.

Notre préoccupation initiale au début de nos recherches était de mesurer les poids relatifs du biais et du bruit dans l'erreur globale. Nous sommes vite arrivés à la conclusion que le bruit contribue souvent plus que le biais à l'erreur globale, et qu'il méritait donc une étude plus approfondie.

Nos premières réflexions sur les différentes composantes du bruit ont été nourries par l'étude d'audits de bruit complexes, dans lesquels un grand nombre d'individus jugeaient de multiples cas. L'étude sur les juges fédéraux du chapitre 6 en est un exemple, et celle sur les dommages punitifs du chapitre 15 en est un autre. Les résultats de ces études fournissent des estimations solides sur le bruit de niveau dans ces jugements. Nous avons vu qu'il peut être important.

Cependant, le design de ces études leur impose une limite dont nous n'avons pas immédiatement mesuré les conséquences. Lorsque chaque participant juge chaque cas, mais ne le fait qu'une fois, il n'est pas possible de dire si l'erreur résiduelle mesurée, que nous avons appelée erreur de pattern, est stable ou transitoire. La prudence des statisticiens les a généralement conduits à choisir la deuxième réponse, et à considérer toute erreur résiduelle comme aléatoire. Autrement dit, l'interprétation par défaut du bruit de pattern dans la plupart des audits de bruit est qu'il consiste entièrement en bruit occasionnel.

Cette interprétation conventionnelle du bruit de pattern – assimilé à une erreur aléatoire – a longtemps limité notre réflexion. Il nous semblait naturel de nous concentrer sur le bruit de niveau : l'existence de différences de niveau moyen, par exemple entre juges cléments et juges sévères, ou entre prévisionnistes optimistes et pessimistes, est un sujet important en soi, d'autant que l'ampleur des différences en question est, on l'a vu, suffisante pour être lourde de conséquences. Nous étions aussi intrigués par les nombreuses études démontrant – en particulier par des méthodes économétriques – l'effet de facteurs transitoires et en principe non pertinents, c'est-à-dire l'existence de bruit occasionnel.

Peu à peu, les données nous ont conduits à adopter un point de vue différent. La variabilité dans les jugements de personnes différentes est largement fonction de quelque chose qui n'est ni un biais général de l'individu, ni un phénomène passager et aléatoire. L'ensemble des réactions d'un individu à une multitude d'aspects des problèmes de jugement crée chez cet individu un pattern spécifique de réactions à chaque problème, aussi unique et aussi stable que sa personnalité. En d'autres termes, l'hypothèse par défaut selon laquelle le bruit de pattern est transitoire ne se justifie pas. Le bruit de pattern stable joue au contraire un rôle essentiel.

Plus précisément, et même s'il convient d'éviter de généraliser à partir d'une sélection d'exemples qui reste limitée, l'ensemble des données dont nous disposons suggère que le bruit de pattern stable est la plus importante des composantes du bruit systémique. Comme nous disposons rarement d'une image complète des différentes composantes de l'erreur dans une même étude, il faut un peu de triangulation pour arriver à cette conclusion indicative. En bref, voici ce que nous savons et ce que nous ne savons pas.

## *La taille des composantes*

Tout d'abord, nous disposons de plusieurs estimations des poids relatifs du bruit de niveau et du bruit de pattern. En règle générale, il semble que le bruit de pattern pèse plus lourd que le bruit de niveau. Pour la compagnie d'assurances du chapitre 2, par exemple, les différences entre les primes moyennes fixées par les différents souscripteurs ne représentaient que 20 % du bruit systémique ; le bruit de pattern représentait donc 80 % du total. S'agissant des juges fédéraux du chapitre 6, le bruit de niveau (les différences de sévérité moyenne) représentait légèrement moins de la moitié du bruit systémique total ; le bruit de pattern était donc plus important. Dans l'expérience portant sur les dommages punitifs, le volume total de

bruit systémique variait largement en fonction de l'échelle utilisée (intention de sanctionner, niveau d'indignation, ou indemnités en dollars), mais la part du bruit de pattern dans ce total était remarquablement constante : elle représentait respectivement 63, 62 et 61 % du bruit systémique total pour les trois échelles. D'autres études que nous évoquerons dans la partie V, en particulier sur les jugements dans le domaine des ressources humaines, confortent cette conclusion indicative.

Le fait que, dans ces études, le bruit de niveau ne soit généralement pas la composante la plus importante du bruit systémique est déjà une conclusion lourde de conséquences. En effet, le bruit de niveau est la seule composante du bruit que les organisations peuvent parfois mesurer sans réaliser d'audits de bruit. Dès lors que les problèmes de jugement sont attribués de manière plus ou moins aléatoire à des professionnels, on peut mesurer les différences entre les moyennes de leurs décisions. Ces différences sont du bruit de niveau.

Tous les étudiants peuvent comparer, par exemple, leurs observations sur le bruit de niveau entre les notations plus ou moins sévères de différents professeurs. De manière plus préoccupante, les études sur l'octroi de brevets[2] font apparaître de grandes différences dans la propension moyenne des inspecteurs à accorder ou à refuser des brevets (avec des répercussions non négligeables sur la fréquence du contentieux dans ce domaine). De même, les travailleurs sociaux employés dans les services de protection de l'enfance varient grandement dans leur propension à retirer les enfants à leur famille pour les placer dans un foyer d'accueil, ce qui n'est bien sûr pas sans conséquences sur les familles, ni, à long terme, sur le bien-être de ces enfants[3].

Ces observations sont étonnantes ; et pourtant, *elles ne portent que sur le bruit de niveau*. Si l'on suppose que, comme dans d'autres domaines, le bruit de pattern est au moins égal (et sans doute supérieur) au bruit de niveau, alors le niveau total de bruit dans l'attribution des brevets ou dans

les décisions de placement d'enfants en danger est *au moins le double* de celui, pourtant déjà choquant, que ces études mettent en évidence. (Notons toutefois qu'il y a nécessairement des exceptions à cette règle indicative. Dans les décisions des juges en matière d'asile, même si l'on peut craindre que le bruit de pattern soit important, il semble clair que le niveau scandaleux de variabilité dans les décisions des juges provient principalement de leurs biais personnels, donc du bruit de niveau[4].)

Une fois le bruit de pattern séparé du bruit de niveau, la seconde étape de l'analyse consiste à analyser ce bruit de pattern en distinguant ses deux composantes. Et comme nous allons le voir, il existe de bonnes raisons de penser que sa composante stable l'emporte sur sa composante occasionnelle.

Pour illustrer notre raisonnement, prenons l'exemple des peines prononcées par les juges fédéraux. En raisonnant par l'absurde, supposons que le bruit de pattern observé provienne entièrement de facteurs transitoires, et constitue donc exclusivement du bruit occasionnel. Cette hypothèse extrême impliquerait que les décisions des juges soient variables, et même erratiques, dans des proportions invraisemblables : la différence moyenne entre deux jugements *de la même affaire par le même juge* (s'ils avaient l'occasion de les juger deux fois) serait d'environ 2,8 ans, à comparer avec la peine moyenne de 7 ans[5]. La variabilité des peines moyennes prononcées par des juges différents est déjà choquante ; mais la même variabilité dans les peines prononcées par le même juge, sur la même affaire, simplement parce que des éléments de contexte ont changé, serait grotesque. Il semble bien plus raisonnable de supposer que les juges diffèrent dans leurs réactions à différents prévenus et à différents crimes, et que ces différences, bien que très personnelles, sont en grande partie stables.

Pour aller au-delà de ces conjectures, et pour quantifier plus précisément les parts respectives du bruit de pattern stable et du bruit

occasionnel, il faut des études dans lesquelles les mêmes juges ont au moins deux occasions de juger les mêmes affaires. Comme nous l'avons mentionné, de telles études sont généralement impossibles sur les questions de jugement complexes, parce qu'il est difficile de garantir que le second jugement soit réellement indépendant du premier. La difficulté tient au fait que les problèmes complexes sont en général mémorables : le juge qui reconnaît un problème déjà traité pourra donc se contenter de réitérer son jugement précédent.

Un groupe de chercheurs de l'Université de Princeton, sous l'égide d'Alexander Todorov, a imaginé des techniques expérimentales ingénieuses pour surmonter cette difficulté[6]. Les chercheurs ont recruté des participants sur Amazon Mechanical Turk, un site où des individus fournissent des services de courte durée – répondre à des questionnaires, par exemple – et sont payés pour le temps qu'ils y consacrent. Dans l'une des expériences, les participants regardaient des images de visages (générées par un programme informatique, mais impossibles à distinguer de visages réels) et leur donnaient une note sur différents critères : trouvaient-ils ces personnes sympathiques ? leur feraient-ils confiance ? Une semaine plus tard – et c'est ce qui rend le dispositif intéressant pour nous – les mêmes participants étaient invités une nouvelle fois à évaluer les mêmes visages sur les mêmes critères. Le grand nombre de questions et le faible enjeu pour les participants permettent d'exclure que ceux-ci aient mémorisé leurs réponses ; et on peut donc considérer les deux occasions de jugement comme indépendantes l'une de l'autre.

Il est raisonnable de s'attendre à trouver moins de consensus entre les participants à cette expérience que chez des professionnels, par exemple de vrais juges traitant des affaires criminelles. Il existe sans doute des visages que presque tout le monde trouve très séduisants et d'autres qui sont universellement repoussants, mais les images générées pour réaliser cette expérience n'appartenaient pas à ces catégories extrêmes, et on pouvait

donc s'attendre à ce que les préférences soient en grande partie personnelles. C'est en effet ce que Todorov et ses collègues constatèrent : le niveau d'accord entre les observateurs était faible. Quand ils analysèrent la variance des réponses à la question « ce visage vous inspire-t-il confiance ? », seuls 18 % provenaient des images elles-mêmes. Les 82 % restants de la variance étaient donc du bruit[7].

Il semble également raisonnable de s'attendre à plus de variabilité transitoire, c'est-à-dire à plus de bruit occasionnel, dans cette étude que dans de véritables jugements professionnels. Il est notoire, en effet, que la qualité des jugements produits par des participants payés pour répondre à des questionnaires en ligne est souvent faible. Malgré cela, la composante la plus importante du bruit était le bruit de pattern stable, suivi par le bruit de niveau, c'est-à-dire les différences entre les observateurs dans les notes moyennes qu'ils donnent à la question de la confiance. Le bruit occasionnel, même s'il restait substantiel, était la plus petite des trois composantes.

Les chercheurs ont abouti aux mêmes conclusions en utilisant d'autres jugements : en demandant aux participants leurs préférences en matière de voitures ou de nourriture, par exemple, mais aussi sur des questions plus proches de ce que nous appelons des jugements professionnels. Ainsi, à notre instigation, les chercheurs utilisèrent leur dispositif expérimental pour reproduire l'étude sur les dommages punitifs que nous avons présentée au chapitre 15. Les participants étaient invités à noter leur « intention de sanctionner » dans dix affaires analogues à celle de « Joan Glover » ; et on leur soumettait à nouveau les mêmes questions une semaine plus tard. Là encore, la composante la plus importante du bruit était le bruit de pattern stable.

En d'autres termes, dans toutes ces études, les participants n'étaient généralement pas d'accord entre eux. Pourtant, ils restaient relativement fidèles à leurs propres jugements d'une occasion de jugement à l'autre, en

dépit du fait que ces jugements n'étaient pas faciles à mémoriser et de l'absence d'enjeu pour eux. Cette « constance sans consensus », pour citer l'expression des chercheurs, témoigne clairement de la présence de bruit de pattern stable.

Un autre résultat qui met en évidence le rôle des patterns stables provient de la grande étude sur les décisions de libération sous caution évoquée au chapitre 10[8]. L'une des analyses que comprend cette remarquable étude est un modèle statistique simulant la manière dont chaque juge utilise les informations disponibles pour décider ou non d'accorder la liberté sous caution. Les auteurs de l'étude ont ainsi construit des « modèles du juge » pour 173 juges différents. Ils ont ensuite soumis à ces 173 juges simulés les 141 833 affaires de leur base de données, ce qui produit plus de 24 millions de jugements au total[9]. À notre demande, les auteurs ont accepté de réanalyser les résultats pour décomposer la variance de ces jugements simulés en trois éléments : la « vraie » variance qui reflète les décisions « justes » (c'est-à-dire moyennes) pour chacune des affaires ; le bruit de niveau créé par les différences dans la propension des juges à accorder la liberté sous caution ; et le bruit de pattern restant.

Cette analyse apporte un éclairage nouveau sur la question qui nous occupe ici, car le bruit de pattern tel qu'il est mesuré par cette méthode est totalement stable. Aucune variabilité aléatoire, donc aucun bruit occasionnel, ne peut être présente dans les « jugements » de *modèles* qui simulent la décision d'un juge. Par construction, ces modèles ne reflètent que les règles de prédiction appliquées par chaque juge avec une stabilité vérifiable.

Le résultat de cette analyse confirme sans équivoque notre hypothèse sur la prépondérance du bruit de pattern stable[10] : celui-ci était presque quatre fois plus important que le bruit de niveau (26 % de la variance totale contre 7 % pour le bruit de niveau). Les patterns idiosyncratiques et stables de jugement qui pouvaient être identifiés contribuaient au bruit de manière

bien plus importante que les différences de sévérité moyenne entre les juges.

Tous ces résultats sont cohérents avec les travaux sur le bruit occasionnel que nous avons examinés au chapitre 7 : l'existence du bruit occasionnel est certes surprenante, voire dérangeante, mais rien n'indique que la variabilité intra-individuelle soit plus importante que les différences interindividuelles[11]. L'élément le plus important du bruit systémique est bien celui que nous avions initialement négligé : le bruit de pattern stable, c'est-à-dire la variabilité entre juges, non pas en moyenne, mais dans leurs jugements de cas particuliers.

Dans la mesure où il existe relativement peu de données sur ce problème, ces conclusions ne peuvent être qu'indicatives, et doivent être formulées avec prudence. Elles nous conduisent néanmoins à réviser notre conception du bruit – et de la manière de s'y attaquer. En principe au moins, le bruit de niveau, c'est-à-dire la présence de différences générales et moyennes entre individus, devrait être un problème relativement facile à mesurer et à résoudre. S'il existe des évaluateurs anormalement « durs », des travailleurs sociaux anormalement « interventionnistes » ou des membres de comités de crédit anormalement « prudents », les organisations qui les emploient peuvent mettre en place des procédures simples pour harmoniser le niveau moyen de leurs jugements. Beaucoup d'universités, par exemple, imposent aux professeurs une distribution prédéterminée des notes au sein de chaque classe, ce qui élimine efficacement le bruit de niveau.

Malheureusement, comme nous le comprenons maintenant, si nous espérons traiter le problème des différences interindividuelles en nous focalisant sur le bruit de niveau, nous passons largement à côté du sujet. La principale source de bruit n'est pas dans les différences de niveau mais dans les interactions : la manière dont différents juges traitent différents prévenus, différents professeurs traitent différents étudiants, différents

travailleurs sociaux traitent différentes familles, différents dirigeants traitent différentes visions du futur. En ce sens, le bruit est essentiellement un sous-produit de notre individualité, de notre unicité : il provient de notre « personnalité de jugement », aussi unique que notre personnalité tout court. Réduire le bruit de niveau ne nuit pas, mais même si cet objectif était atteint, la plus grande partie du problème resterait sans réponse.

## *Expliquer l'erreur*

Comme vous l'aurez constaté, nous sommes intarissables sur le sujet du bruit. Pourtant, celui-ci est presque totalement absent du débat public, et même des échanges entre professionnels ou entre spécialistes du jugement et de l'erreur. Malgré les preuves de sa présence et la multiplicité des mécanismes qui le produisent, le bruit n'est que rarement mentionné comme un facteur important du jugement. Comment cela est-il possible ? Pourquoi n'invoquons-nous jamais le bruit pour expliquer les erreurs de jugement, tandis que nous en accusons couramment les biais ? Pourquoi tant de silence sur le bruit ?

La clé de cette énigme est que si la moyenne des erreurs (le biais) et leur variabilité (le bruit) jouent un rôle équivalent dans l'équation de l'erreur, la façon dont nous les appréhendons psychologiquement est radicalement différente. Notre manière de trouver du sens au monde qui nous entoure se satisfait du biais, mais rend presque impossible de percevoir le rôle du bruit.

Précédemment dans ce livre, nous avons noté qu'il nous était facile de donner un sens aux événements *a posteriori*, même si nous n'avons pas pu les prévoir avant qu'ils ne se produisent. Dans la vallée des Choses Normales, les événements ne sont jamais surprenants, et ils sont toujours faciles à expliquer.

Ce qui est vrai des événements l'est aussi des jugements. Comme d'autres événements, les jugements et les décisions se produisent pour la plupart dans la vallée des Choses Normales : en général, ils ne nous surprennent pas. C'est le cas, tout d'abord, parce que les jugements qui produisent des résultats satisfaisants sont considérés comme normaux. Personne ne songe à les remettre en cause. Quand le tireur choisi pour un lancer franc marque le panier, quand le chirurgien réussit son opération, ou quand l'entreprise sur laquelle un investisseur a misé prospère, nous supposons que les raisons qui ont conduit les décideurs à leurs choix devaient être les bonnes. Après tout, ils ont vu juste ! Comme n'importe quelle histoire qui ne constitue pas une surprise, une histoire de succès se suffit à elle-même : une fois son dénouement connu, elle est sa propre explication.

Il arrive toutefois que nous ressentions un besoin d'explication. C'est le cas quand les résultats sont anormaux : soit parce que l'échec est patent, soit parce que le succès nous étonne (par exemple, quand un pari manifestement risqué réussit de manière spectaculaire).

Dans ce type de situation, les explications font le plus souvent appel à la personne des décideurs : on invoque, respectivement, leurs erreurs et leur incompétence, ou leur talent et leur flair exceptionnels. Ces explications sont beaucoup plus fréquentes qu'elles ne méritent de l'être. Un biais psychologique solidement établi, l'*erreur d'attribution fondamentale*, nous conduit à attribuer aux personnes des actions et des résultats qui s'expliquent en réalité par la chance ou par des circonstances objectives. Un autre biais, le *biais rétrospectif*, déforme nos jugements de telle sorte que des résultats qu'il était impossible d'anticiper apparaissent *a posteriori* facilement prévisibles. C'est ainsi qu'un pari incertain dont le résultat doit beaucoup à la chance se mue en coup de génie ou en coup de folie de son auteur – une fois que le résultat est connu.

De manière générale, il est donc très facile de trouver des explications aux erreurs de jugement. De la même manière que nous trouvons aisément des causes aux événements une fois que ceux-ci se sont produits, nous en trouvons, plus facilement encore, aux jugements, une fois que nous les avons observés. Nous pouvons généralement prêter à l'auteur d'un jugement des motivations qui l'expliquent. Si cela ne suffit pas, par exemple parce que quelqu'un a manifestement nui à ses propres intérêts, il suffit d'invoquer son incompétence.

À cet arsenal explicatif s'ajoute une arme nouvelle, dont l'usage pour analyser les mauvais jugements est devenu fréquent depuis quelques décennies : les biais psychologiques. Un corpus abondant de recherches en psychologie et en économie comportementale a identifié une longue liste de biais psychologiques : l'erreur de prévision, l'excès de confiance, l'aversion à la perte, l'effet de dotation, le biais du *statu quo*, le biais pour le présent, et beaucoup d'autres, sans oublier, bien sûr, les biais favorables ou défavorables à certaines catégories de personnes. On en sait aujourd'hui beaucoup sur les circonstances dans lesquelles chacun de ces biais a des chances d'influencer les jugements et les décisions, et sur ce qui peut permettre à un observateur de la prise de décision de reconnaître en temps réel un raisonnement biaisé.

Il est tout à fait légitime d'invoquer un biais psychologique comme explication causale d'une erreur de jugement, à condition que le rôle du biais en question puisse être prédit à l'avance ou détecté en temps réel. Si le biais psychologique n'est identifié qu'une fois le résultat connu, il peut encore fournir une hypothèse d'explication, au moins à titre provisoire, à condition que cette explication constitue aussi une prédiction pour l'avenir. Quand, par exemple, un comité de sélection, contre toute attente, met à l'écart une excellente candidate à un poste, on peut formuler l'hypothèse plus générale qu'un biais de genre est à l'œuvre, hypothèse que les futures nominations décidées par le même comité confirmeront ou infirmeront.

Considérons, à l'inverse, une explication causale qui ne s'applique qu'à un seul événement et n'a pas de valeur prédictive : « Ils ont échoué dans cette entreprise parce qu'ils ont péché par excès de confiance. » L'affirmation est totalement creuse, mais elle donne une illusion de compréhension qui peut être assez satisfaisante. Phil Rosenzweig, dans ses travaux sur les stratégies d'entreprise[12], a dénoncé avec vigueur ces explications *a posteriori* fondées sur les biais, qui sont fréquentes dans l'interprétation des résultats. Leur popularité témoigne de la force du besoin de trouver des histoires causales qui donnent un sens à ce que nous vivons.

## *Le bruit est statistique*

Comme nous l'avons remarqué au chapitre 12, notre mode de pensée normal est la pensée causale. Nous prêtons naturellement attention aux cas particuliers, en suivant et en inventant des séquences cohérentes de causes et d'effets qui expliquent ces cas. Les échecs y sont souvent attribués à des erreurs, et ces erreurs à des biais. Rien de tout cela n'est difficile ; et cette facilité à expliquer les erreurs de jugement ne laisse guère de place au bruit.

L'inaudibilité du bruit est une conséquence directe de la pensée causale. Le bruit est un phénomène intrinsèquement statistique : il ne devient perceptible que lorsque nous réfléchissons sur un ensemble de jugements similaires, en utilisant la pensée statistique plutôt que la pensée causale. Dès lors, bien sûr, il devient impossible de le manquer. Qu'on l'analyse *a posteriori* – par exemple, sous forme de variabilité dans les peines de prison ou les primes d'assurance – ou *a priori* – par exemple en mesurant l'incertitude dans les prévisions –, le bruit devient assourdissant dès qu'on considère un ensemble, plutôt qu'un seul cas. L'analogie du stand de tir s'applique ici encore : un tir qui touche la cible ou qui la manque peut aisément s'expliquer par le talent du tireur ou par sa maladresse ; mais ce

n’est que lorsqu’on observe une série de tirs que la dispersion des impacts sur la cible devient apparente. Quand on utilise la pensée causale, le bruit n’est nulle part ; quand on adopte la pensée statistique, il est partout.

Hélas, la pensée statistique n’est pas naturelle. Nous trouvons sans aucun effort des causes aux événements que nous observons, mais pour penser statistiquement, il faut un apprentissage initial et un effort de chaque instant. Les causes sont aisées, les statistiques difficiles.

---

Le bruit systémique est une source importante d’erreur ; et de nombreux indices suggèrent que sa composante la plus importante est le bruit de pattern stable. Pourtant, nous n’y prêtons généralement guère attention. En grande partie parce que notre mode naturel de pensée est causal, il existe un déséquilibre marqué dans l’attention que nous portons au biais – qui est une cause identifiable d’erreur – et au bruit – qui ne peut s’appréhender qu’en adoptant la pensée statistique.

Si vous avez suivi un cours d’introduction à la psychologie de la perception, vous vous rappelez peut-être ces illustrations où l’on voit une silhouette se détacher sur un arrière-plan flou. Notre attention se fixe durablement sur le premier plan, même si l’immensité du fond la fait paraître minuscule. Le contraste entre la silhouette et l’arrière-plan est une bonne métaphore de nos intuitions sur le biais et le bruit : le biais est la silhouette qui capte le regard, le bruit est l’arrière-plan auquel nous ne prêtons pas attention. C’est ainsi que nous restons dans une large mesure inconscients d’une faille importante de notre jugement.

## À propos des sources du bruit

« Nous voyons facilement les différences dans le niveau moyen des jugements, mais quelle est l'ampleur du bruit de pattern que nous ne voyons pas ? »
« Vous dites que ce jugement a été causé par des biais, mais diriez-vous la même chose si le résultat avait été différent ? Et pouvez-vous dire s'il y avait du bruit ? »
« Nous nous sommes attachés à réduire les biais que nous remarquons, et nous avons raison ! Mais attachons-nous aussi à réduire le bruit auquel nous ne songeons pas aussi souvent. »

PARTIE V

# Améliorer les jugements

Comment une organisation peut-elle améliorer les jugements professionnels qui y sont produits ? Comment peut-elle, en particulier, en diminuer le bruit ? S'il vous fallait répondre à ces questions, que feriez-vous ?

La première étape est la prise de conscience : comprendre qu'il y a du bruit dans les jugements professionnels, et que ce bruit est un problème qui mérite l'attention des dirigeants. C'est l'objet d'un audit de bruit (dont l'annexe A donne un mode d'emploi détaillé). Dans un audit de bruit, un grand nombre d'individus sont appelés à se prononcer sur les mêmes problèmes. On mesure ainsi le bruit systémique, qui est la variabilité de leurs jugements.

Dans certains cas, bien sûr, cette variabilité pourra être imputée à l'incompétence : une partie des individus appelés à formuler des jugements peuvent tout simplement ne pas savoir ce qu'ils font. Le problème peut être un déficit général de compétences ou de motivation, affectant donc tous les cas, ou une absence de préparation à certains types de cas qui soulèvent des difficultés particulières. Quand de tels déficits de compétences existent, il va sans dire que la priorité est d'y remédier. Mais, comme nous l'avons vu, l'incompétence de quelques-uns n'est pas la seule cause du bruit dans les

jugements de tous. Il peut y avoir une grande quantité de bruit même dans les jugements de professionnels qualifiés, motivés et bien formés.

Si l'audit de bruit révèle un niveau de bruit systémique intolérable, l'une des options à envisager est de remplacer le jugement humain par des règles ou par des algorithmes. Comme nous l'avons vu, cette solution radicale permet d'éliminer entièrement le bruit. Mais elle soulève d'autres difficultés (comme nous le verrons dans la partie VI). Même les partisans les plus enthousiastes de l'intelligence artificielle ne suggèrent pas de la substituer au jugement humain dans tous les domaines, ni aujourd'hui, ni dans un avenir prévisible. L'amélioration du jugement humain – plutôt que son remplacement – est donc une tâche plus urgente que jamais. Elle constitue le sujet de cette partie du livre.

Une première idée qui vient naturellement à l'esprit quand on parle d'améliorer le jugement est bien sûr de sélectionner les meilleurs juges possibles. Au stand de tir, certains tireurs ont l'œil plus aiguisé et le geste plus sûr. De même, dans tout jugement professionnel, certains individus font moins d'erreurs, parce qu'ils sont moins sujets à la fois au bruit et au biais. La question devient alors de savoir comment identifier ces individus supérieurs. Quelquefois, c'est assez simple : si vous voulez résoudre un problème d'échecs, vous vous adresserez à un grand maître plutôt qu'aux auteurs de ce livre. Mais dans la plupart des cas, ce qui caractérise le très bon juge est plus difficile à cerner. Ce sera le sujet du chapitre 18.

Nous examinerons ensuite les différentes manières de réduire les erreurs de jugement. Comme nous l'avons vu, les biais psychologiques sont impliqués à la fois dans le biais statistique et dans le bruit. Au chapitre 19, nous passerons en revue les différents types de démarches qui visent à réduire ces biais psychologiques, avec des succès divers. Pour compléter ces stratégies de réduction des biais, nous en proposerons une qui nous semble prometteuse et qui, à notre connaissance, n'a pas été systématiquement explorée : désigner un *observateur de la décision* qui

diagnostiquera, en temps réel, la présence d'un ou plusieurs biais connus dans un processus de décision.

Cependant, ce n'est pas la réduction du biais qui constitue le cœur de cette partie du livre, mais bien la lutte contre le bruit. Nous introduirons l'idée d'*hygiène de la décision* pour réduire le bruit dans les jugements humains : comme l'hygiène sanitaire, l'hygiène de la décision est une démarche fondamentalement préventive, qui vise à limiter le bruit avant qu'il n'ait produit ses effets. Nous l'illustrerons par cinq chapitres présentant des études de cas dans autant de domaines différents. Dans chacun de ces domaines, nous analyserons la prévalence du bruit et rappellerons les dommages qu'il peut occasionner. Nous examinerons aussi les stratégies qui ont été utilisées pour réduire le bruit, avec plus ou moins de succès.

Dans chaque domaine, bien sûr, de multiples approches ont été utilisées, mais pour la commodité de l'exposé, chaque chapitre mettra l'accent sur une technique particulière d'hygiène de la décision. Dans le chapitre 20, nous commencerons par le cas de la police scientifique, qui illustre l'importance du *séquencement de l'information*, permettant d'éviter que des informations inutiles ne parasitent le processus de jugement. Dans le chapitre 21, nous parlerons des prévisions en tous genres, qui illustrent une stratégie de réduction du bruit particulièrement polyvalente : l'*agrégation de jugements indépendants*. Le chapitre 22 nous emmènera dans le monde médical, où nous retrouverons une stratégie de réduction du bruit déjà mentionnée à propos du jugement pénal : les *directives*. Dans le chapitre 23, nous aborderons un problème bien connu des entreprises : l'évaluation de la performance, qui met en évidence l'importance d'utiliser une *échelle partagée calibrée par une vision externe*. Le chapitre 24, enfin, explorera le sujet connexe mais distinct du recrutement, au sujet duquel de très nombreuses recherches depuis plus d'une centaine d'années démontrent l'intérêt de la *structuration des jugements complexes*.

Les enseignements du recrutement peuvent être étendus à d'autres problèmes de choix entre options, en s'appuyant sur une idée simple : « les options sont comme des candidats ». Dans le chapitre 25, nous proposerons donc une approche générale de l'évaluation des options, le *protocole d'évaluations intermédiaires*. Le protocole d'évaluations intermédiaires combine la structuration et d'autres techniques d'hygiène de la décision, et peut s'appliquer aux décisions récurrentes comme aux décisions singulières.

Une dernière observation s'impose en préambule de cette partie. Il serait intéressant de pouvoir quantifier les gains attendus de chaque technique d'hygiène de la décision dans différents contextes et de pouvoir choisir les plus efficaces. Par exemple, quand l'information est séquencée avec soin, dans quelle mesure le bruit est-il réduit ? Si l'on choisit d'agréger des jugements multiples, en pratique, combien faut-il de jugements ? Dans quels domaines la structuration des jugements apporte-t-elle plus d'avantages que d'inconvénients ?

Comme le sujet du bruit n'a suscité jusqu'ici qu'un intérêt limité, toutes ces questions restent ouvertes, et y répondre de manière systématique constitue un champ de recherche prometteur. En pratique, les avantages de chaque technique d'hygiène de la décision dépendront du contexte dans lequel chacune est utilisée. La mise en place de directives, par exemple, produira quelquefois des résultats impressionnants (comme nous le verrons dans certains diagnostics médicaux) ; mais dans d'autres types de jugements, son impact sera modeste (comme nous le verrons sur d'autres exemples dans le même chapitre). À ce stade, le choix des techniques d'hygiène de la décision les plus adaptées à un contexte donné est donc une question empirique. La réponse doit prendre en compte les bénéfices probables de chaque technique, mais aussi ses coûts : un arbitrage que nous aborderons dans la partie VI.

CHAPITRE 18

# De meilleurs juges pour de meilleurs jugements

Jusqu'à présent, quand nous avons parlé des jugements humains et des juges qui les produisent, nous n'avons pas fait de distinctions entre ces juges. Or il semble évident que dans tout travail de jugement, certains individus sont meilleurs que d'autres. Même un jugement produit par agrégation (par la « sagesse des foules ») sera sans doute meilleur si la foule en question se compose d'individus plus compétents[1]. Quiconque souhaite obtenir de meilleurs jugements doit donc se demander comment identifier de meilleurs juges.

Comme nous allons le voir, ces « meilleurs juges », moins sujets au bruit et au biais, se distinguent de trois manières : la compétence technique, l'intelligence, et le style cognitif. Il est évident que la qualité de vos jugements dépend de ce que vous savez dans votre spécialité ; et vous ne serez pas surpris qu'elle dépende aussi de vos aptitudes intellectuelles. Plus étonnant est le rôle que joue le style cognitif, c'est-à-dire votre *manière* de réfléchir. Les bons juges ne sont pas seulement qualifiés et intelligents : ce sont aussi des individus « activement ouverts d'esprit », capables de tirer des enseignements d'informations nouvelles.

## *Les experts et les experts d'estime*

Il est presque tautologique de dire que la qualité des jugements est fonction de la compétence des juges. Ainsi, un radiologue plus compétent a plus de chances de diagnostiquer correctement une pneumonie, et en matière de prévision géopolitique, il existe des « superprévisionnistes » qui voient généralement plus juste que leurs pairs. Si vous réunissez un groupe d'avocats spécialistes du droit des contrats, leurs prévisions quant à l'issue d'un procès normalement complexe portant sur un litige de ce type seront probablement similaires entre elles, et proches de la réalité. En d'autres termes, ils seront moins sujets au bruit, et n'auront pas non plus de biais statistique important.

Dans ces exemples, les individus en question peuvent être considérés comme de véritables experts. Leur supériorité sur leurs pairs peut être facilement confirmée : comme leurs jugements sont prédictifs et vérifiables, on peut les comparer à des résultats réels. En principe au moins, il devrait être possible de choisir un médecin, un prévisionniste ou un avocat en fonction du nombre de fois où il a vu juste dans ses diagnostics, dans ses prévisions économiques, ou dans ses pronostics sur l'issue d'un procès.

Pour des raisons évidentes, cette approche est difficile à mettre en pratique : nous ne vous recommandons pas de soumettre votre médecin de famille à un test d'aptitudes. Surtout, comme nous l'avons vu, de nombreux jugements ne sont *pas* vérifiables. Ni les jugements des souscripteurs du chapitre 2, ni ceux des juges pénaux du chapitre 6 ne peuvent être comparés à une valeur vraie. De même, la dégustation de vins, la notation de dissertations, la critique de livres et de films, et maints autres sujets donnent lieu à des jugements qui ne sont pas objectivement vérifiables.

Dans ce type de situation, et à l'exception des cas où un jugement est manifestement et grossièrement faux, il est impossible d'en évaluer la qualité. Puisqu'on ne peut pas connaître (ni même parfois définir de

manière incontestable) la valeur vraie dont ces jugements tentent de s'approcher, comment pourrait-on identifier les « bons » jugements dans ces domaines ? Cela n'empêche pas que certains professionnels y soient considérés comme des experts. La confiance que nous avons dans leur jugement est entièrement fondée sur l'estime que leur témoignent leurs pairs. Nous les appelons des *experts d'estime*.

L'expression *expert d'estime* n'entend pas être péjorative. Ce n'est pas critiquer ces experts que d'observer que la justesse de leurs jugements n'est pas objectivement mesurable. C'est un simple constat ; et il est valable dans beaucoup de domaines. Les auteurs de ce livre, comme la plupart des professeurs, des universitaires et des consultants en management, sont des experts d'estime. Leur crédibilité repose sur le respect que leur accordent leurs étudiants, leurs pairs ou leurs clients. Dans ces domaines, et dans beaucoup d'autres, les jugements d'un professionnel ne peuvent être comparés qu'à d'autres jugements.

En l'absence de valeurs vraies permettant de déterminer qui a tort et qui a raison, on se trouve souvent dans une situation paradoxale, où l'on a de l'estime pour des experts qui ne sont pas d'accord entre eux. Imaginez, par exemple, un plateau de télévision où trois analystes politiques exposent des opinions très différentes sur les raisons d'une grave crise diplomatique et sur son issue prévisible. (Ce n'est pas rare : le débat manquerait de piquant si les experts étaient unanimes.) Chacun des analystes croit, bien sûr, qu'il est possible de faire une analyse et une prédiction correctes à propos de cette crise. Chacun, également, sans pour autant se croire infaillible, est convaincu d'être, sur ce plateau, celui qui s'en rapproche le plus. En écoutant les politistes présenter leurs points de vue, vous les trouvez tous les trois remarquables, et leurs arguments également convaincants. Vous n'avez aucun moyen de savoir lequel a raison. Vous ne le saurez peut-être même pas après coup, car leurs prédictions ne sont pas forcément formulées de manière suffisamment précise pour être vérifiables. Vous savez que

certains d’entre eux doivent nécessairement se tromper, puisqu’ils sont en désaccord. Et pourtant, vous avez de l’estime pour l’expertise de chacun.

Ou prenons un autre type d’experts, qui ne font pas de prédictions du tout. Trois spécialistes de philosophie morale, tous diplômés des meilleures institutions et auteurs d’ouvrages remarqués, sont réunis autour d’une table. Le premier est un disciple d’Emmanuel Kant, le deuxième s’inspire de l’utilitarisme de Jeremy Bentham, et le troisième ne jure que par Aristote. Évidemment, ils ne sont pas du tout d’accord. Que le débat porte sur la légitimité du mensonge dans certaines circonstances, sur les droits des animaux ou sur la finalité de la sanction pénale, leurs raisonnements et leurs conclusions sont radicalement différents. Vous les écoutez avec attention. Vous admirez la précision et la clarté de leur pensée. Sur chaque sujet, vous aurez peut-être tendance à être d’accord avec l’un des trois, mais vous les respecterez tous.

Comment est-ce possible ? Plus généralement, comment des individus qui sont eux-mêmes estimés pour la qualité de leur jugement décident-ils d’accorder leur estime à quelqu’un d’autre et de le considérer comme un expert, en l’absence de données permettant de valider objectivement son expertise ? Qu’est-ce qui fait l’expert d’estime ?

Une première réponse est qu’il existe toujours, dans une profession donnée, un ensemble de normes et de pratiques généralement admises. Beaucoup d’experts suivent une formation initiale et reçoivent une accréditation accordée par une instance représentant leur profession. Par la suite, ils continuent à se former au sein des organisations qui les emploient et les supervisent. Les médecins qui font leur internat ou les avocats débutants qui apprennent leur métier auprès d’un associé ne pratiquent pas seulement des techniques : ils développent aussi leur jugement, en apprenant à faire usage des méthodes de leur profession et à se conformer à ses normes.

Ces méthodes et ces normes partagées fournissent aux professionnels de solides indications sur le type d'informations à prendre en compte pour réaliser un jugement ou pour le justifier *a posteriori*. Dans notre compagnie d'assurances, par exemple, les experts en sinistres étaient tout à fait d'accord sur la liste des facteurs importants à prendre en compte dans l'évaluation d'un dossier.

Ce consensus sur la méthode n'empêchait pas, bien sûr, que les experts soient en complet désaccord sur beaucoup d'estimations. La méthode n'est pas une recette qu'ils suivent à la lettre. Elle laisse au contraire une large place à l'interprétation. Ce que les experts produisent est bien un jugement, pas un calcul qui pourrait être mécaniquement reproduit. C'est ce qui rend le bruit inévitable. Même des professionnels formés de manière identique et qui s'entendent sur les méthodes et normes professionnelles en vigueur s'écartent inéluctablement les uns des autres dans l'application qu'ils en font.

Il ne suffit pas, pour être un expert d'estime, de maîtriser les normes professionnelles : il faut aussi de l'expérience. Si votre spécialité est le jeu d'échecs, le piano de concert ou le lancer du javelot, vous avez peut-être été un jeune prodige : votre niveau de performance était validé par vos résultats. Mais les experts en sinistres, les spécialistes des empreintes digitales ou les juges d'assises ont généralement besoin de longues années d'expérience pour être crédibles. Il n'y a pas de jeunes prodiges chez les souscripteurs d'assurances.

Les experts d'estime se distinguent enfin par leur capacité à produire et expliquer leurs jugements avec aplomb. Leur expérience leur permet de reconnaître des situations, de raisonner par analogie, de former et de confirmer rapidement des hypothèses. Ainsi, ils manient facilement les faits pour les assembler en un récit cohérent, qui leur permet d'avoir confiance dans leur jugement. Or nous avons tendance à nous fier davantage aux gens sûrs d'eux qu'à ceux qui manifestent des doutes. C'est l'*heuristique de la*

*confiance* : dans un groupe, les individus confiants ont plus de poids que les autres, même si leur confiance en eux n'est pas fondée sur une supériorité réelle[2]. Si les experts d'estime nous inspirent confiance, c'est en partie parce qu'ils ont confiance en eux.

## *L'intelligence*

La formation, l'expérience et la confiance en soi permettent donc aux experts d'estime d'inspirer le respect. Mais ces qualités ne garantissent nullement que leurs jugements seront corrects. Y a-t-il d'autres indicateurs qui permettent d'identifier les experts capables de produire de bons jugements ?

À cette question, il existe de solides raisons de proposer une réponse qui tient en un mot : l'intelligence. Une intelligence supérieure a de fortes chances d'être associée à de meilleurs jugements, parce que l'intelligence est corrélée avec la performance dans presque tous les domaines. En particulier, et toutes choses égales par ailleurs, l'intelligence est associée non seulement à la réussite scolaire, mais aussi à une meilleure performance au travail[3].

Plutôt que d'intelligence ou de quotient intellectuel, on préfère aujourd'hui généralement parler d'aptitudes mentales générales (*general mental ability* ou GMA). Cette précaution terminologique ne suffit pas pour autant à prévenir des malentendus et des débats souvent houleux. En particulier, on croit souvent, à tort, que l'intelligence est innée[4], alors que les tests mesurent des capacités « développées » : celles-ci sont certes en partie fonction de caractères héréditaires, mais elles dépendent aussi de l'environnement, notamment de l'accès à l'éducation. Par ailleurs, beaucoup s'inquiètent aujourd'hui des effets sur certains groupes sociaux

d’une sélection fondée sur la GMA, et contestent par conséquent la légitimité de l’utilisation de tests d’aptitude à des fins de sélection.

Ces questions sur l’utilisation des tests doivent être distinguées du sujet qui nous occupe ici : leur valeur prédictive bien réelle. Depuis que l’armée américaine a commencé à utiliser des tests d’aptitudes mentales, il y a plus d’un siècle, des milliers d’études ont permis de vérifier le lien statistique entre les notes obtenues à ce type de tests cognitifs et la performance ultérieure. Le message qui se dégage de cette masse de recherches est sans ambiguïté. Comme le résumait en 2004 une revue de la littérature scientifique sur le sujet, « la GMA prédit à la fois le niveau professionnel atteint et la performance dans le métier choisi, et sa valeur prédictive est supérieure à celle de toute autre aptitude, trait ou disposition, et même à celle de l’expérience professionnelle[5] ».

Bien sûr, d’autres capacités cognitives comptent aussi (nous y reviendrons). Et de nombreux traits de personnalité sont aussi de bons prédicteurs de la réussite. C’est notamment le cas du caractère consciencieux et du *grit*, défini comme un mélange d’ardeur et de persévérance dans la poursuite d’objectifs de long terme (ce qu’on appelle familièrement en français la « niaque »[6]). Il est vrai, enfin, qu’il existe plusieurs formes d’intelligence, comme l’intelligence pratique et la créativité, et que toutes ne sont pas mesurées par les tests de GMA. En particulier, les psychologues et les neuroscientifiques distinguent l’intelligence cristallisée, qui est la capacité de résoudre des problèmes en mobilisant des connaissances (par exemple arithmétiques), de l’intelligence fluide[7], qui est l’aptitude à résoudre des problèmes nouveaux.

Pourtant, pour rudimentaire et limitée qu’elle soit, la GMA, telle que la mesurent des tests standardisés à base de problèmes verbaux, quantitatifs et spatiaux, reste de loin le meilleur prédicteur d’un certain nombre de résultats importants. Comme le note l’étude de 2004 mentionnée plus haut, la puissance prédictive de la GMA est « plus grande que ce que l’on

observe le plus souvent dans la recherche en psychologie[8] ». Le lien statistique entre aptitudes mentales et réussite au travail augmente, assez logiquement, avec la complexité du travail en question : il est plus utile d'être intelligent si vous êtes astrophysicien que si votre travail consiste à répéter des tâches simples. Pour les emplois très complexes, les corrélations observées entre les résultats aux tests standardisés de GMA et la performance au travail sont autour de 0,50 (PC = 67 %)[9]. Comme nous l'avons vu, en sciences sociales, une corrélation de 0,50 correspond à une puissance prédictive inhabituellement élevée[10].

On objecte souvent que les mesures d'intelligence perdent de leur pertinence quand il est question de professions intellectuelles supérieures, dans la mesure où les individus qui produisent ce type de jugement ont tous des aptitudes mentales élevées. Les médecins, les juges, ou les souscripteurs d'assurances dont nous avons évoqué ici les jugements professionnels ont en effet des niveaux d'études très supérieurs à la moyenne, et obtiendraient sans doute aussi des scores élevés à des tests d'aptitudes mentales générales. Il semble donc raisonnable de penser que la valeur exacte de ce score ne change pas grand-chose pour eux. En d'autres termes, on pourrait penser qu'une intelligence supérieure constitue le « ticket d'entrée » dans ces professions, mais pas une source de différenciation entre ceux qui y appartiennent.

Cette théorie, bien que largement répandue, est erronée. Il ne fait certes aucun doute que des aptitudes mentales élevées conditionnent l'accès à certains emplois. On peut le vérifier en comparant l'éventail des scores de GMA dans différents métiers : on trouve des individus avec une GMA élevée dans des métiers peu qualifiés, mais presque aucun individu ayant une GMA inférieure à la moyenne[11] chez les avocats, les pharmaciens ou les ingénieurs. Cependant, cette observation ne rend pas compte des différences de réussite au sein même de ces professions. Même parmi les 1 % des individus dont les aptitudes cognitives (évaluées à l'âge de 13 ans)

sont les plus élevées, les résultats exceptionnels sont corrélés avec la GMA[12]. Si l'on divise ce « top 1 % » en quatre sous-groupes de taille égale, ceux qui sont dans le quartile supérieur obtiennent des résultats très différents de ceux qui sont dans le quartile inférieur : ils ont une probabilité deux à trois fois supérieure d'obtenir un doctorat, de publier un livre ou de déposer un brevet. Autrement dit, non seulement les écarts de GMA sont importants quand on compare le 99e percentile au 80e ou au 50e, mais ils comptent encore – et beaucoup ! – entre le 99,88e percentile et le 99,13e.

Une étude de 2013 fournit une autre illustration frappante du lien entre aptitudes générales et résultats. L'étude s'est intéressée aux P-DG des cinq cents plus grandes entreprises américaines (les *Fortune 500*[13]) et aux 424 milliardaires américains (le top 0,0001 % des plus riches). L'étude confirme, comme on pouvait s'y attendre, que ces groupes de l'hyperélite sont composés d'individus à hautes aptitudes intellectuelles. Mais l'étude montre aussi qu'*à l'intérieur* de ces groupes, le niveau d'éducation et d'aptitudes mentales est corrélé au niveau de rémunération (pour les P-DG) et de fortune (pour les milliardaires). On notera au passage que le cliché de l'entrepreneur milliardaire qui a abandonné l'université, comme l'ont fait Steve Jobs, Bill Gates ou Mark Zuckerberg, est l'arbre qui cache la forêt : 88 % des milliardaires ont un diplôme universitaire, alors que ce n'est le cas que pour un tiers environ des adultes américains.

La conclusion s'impose : le niveau d'aptitudes mentales générales contribue largement à la réussite dans les professions qui nécessitent du jugement. Cela reste vrai quand on compare des individus qui ont tous, dans l'absolu, un niveau élevé d'aptitudes. L'idée qu'il existerait un seuil au-delà duquel l'intelligence cesse de faire la différence est démentie par les faits.

Cette observation en suggère une autre. Quand les jugements professionnels sont invérifiables, c'est-à-dire quand la valeur vraie dont ils cherchent à s'approcher n'est pas connaissable, les jugements produits par des individus ayant un niveau élevé d'aptitudes mentales auront une plus

forte probabilité de se rapprocher de cette valeur vraie. Par conséquent, si vous devez choisir à qui vous fier pour produire des jugements, il est logique de sélectionner les individus ayant les aptitudes mentales les plus élevées.

Ce raisonnement se heurte toutefois à une limite importante. Sauf à faire passer des tests standardisés à tout le monde, il vous faudra deviner par vous-même qui sont ces individus aux aptitudes mentales élevées. Or si les aptitudes mentales générales méritent d'être appelées « générales », c'est parce qu'elles contribuent à améliorer les performances dans de nombreux exercices… y compris celui qui consiste à convaincre d'autres personnes que l'on a raison ! Les individus dotés d'aptitudes mentales élevées ont certes de fortes probabilités de produire de bons jugements, mais ils ont aussi de fortes probabilités d'impressionner leurs pairs et de gagner leur confiance. Autrement dit, si le domaine dans lequel ils exercent leur jugement ne permet pas qu'on mesure leurs résultats réels, ils ont toutes les chances de devenir des experts d'estime. Les scores de GMA des astrologues du Moyen Âge devaient être parmi les plus élevés de leur temps.

En somme, il peut être raisonnable de placer votre confiance dans des gens qui ont l'air intelligent et qui justifient leurs jugements de manière convaincante, mais cette stratégie n'est pas suffisante, et elle peut même s'avérer contre-productive. À quels autres traits, alors, peut-on identifier les vrais experts ?

## *Le style cognitif*

Indépendamment des aptitudes mentales, les individus diffèrent par leur *style cognitif*, c'est-à-dire par leur approche des tâches de jugement. De nombreux tests ont été développés pour rendre compte des styles cognitifs.

La plupart des mesures qu'ils produisent sont corrélées avec la GMA (et corrélées entre elles), mais ce qu'elles mesurent est légèrement différent.

Un de ces tests est le *test de réflexion cognitive* (*cognitive reflection test* ou CRT), rendu célèbre par la question de la balle et de la batte : « Une balle et une batte coûtent 1,10 dollar au total. La batte coûte 1 dollar de plus que la balle. Quel est le prix de la balle ? » Autre exemple de question qui mesure la réflexion cognitive : « Si vous participez à une course et que vous dépassez la personne qui est à la deuxième place, à quelle place êtes-vous[14] ? » À ces deux questions, une réponse vient spontanément à l'esprit : « 10 *cents* » pour la question de la balle et de la batte, « en première position » pour la question de la course. Mais ces réponses intuitives sont incorrectes, comme quelques instants supplémentaires de réflexion permettent de le constater.

Le CRT mesure la capacité qu'ont les individus à s'engager dans cette réflexion supplémentaire pour vérifier leur première intuition, constater qu'elle est fausse, et lui substituer la réponse exacte. Cette capacité semble jouer un rôle dans la formation des jugements et des croyances : les personnes qui obtiennent des résultats plus faibles au CRT sont plus nombreuses à croire aux fantômes, à l'astrologie, et à la perception extrasensorielle[15]. Elles ont plus de chances de se laisser berner par des *fake news* manifestement inexactes[16]. On a même pu mesurer que les personnes au CRT plus faible passent un temps plus important à utiliser leurs smartphones[17].

Le CRT peut être considéré comme une mesure de la propension à utiliser les processus de la pensée réfléchie plutôt que ceux de la pensée impulsive[18]. Pour le dire plus simplement, certaines personnes aiment réfléchir attentivement, tandis que d'autres, face au même problème, ont tendance à faire confiance à leur intuition première. Dans notre terminologie, le CRT est une mesure de la propension à recourir à la pensée lente du Système 2 plutôt qu'à la pensée rapide du Système 1.

D’autres tests ont été développés pour mesurer cette propension (et tous ces tests sont, bien sûr, corrélés entre eux). Par exemple, l’échelle du besoin de cognition (*need for cognition*, ou NFC) demande aux gens quelle est leur appétence pour la réflexion sur des problèmes ardus[19]. Vous obtiendrez une note élevée sur cette échelle si vous exprimez votre accord avec des propositions comme : « J’ai tendance à me donner des objectifs qui ne peuvent être atteints qu’au prix d’un effort mental considérable. » Vous aurez un score plus faible si vous souscrivez à l’affirmation : « Réfléchir, ça ne m’amuse pas du tout. » Les individus qui obtiennent un score élevé sur l’échelle du besoin de cognition ont tendance à mieux résister aux biais cognitifs[20]. Certaines associations plus étranges ont aussi été signalées : si vous détestez qu’une critique de film contienne des « spoilers » révélant des détails de l’intrigue, vous avez probablement un fort besoin de cognition ; les gens qui n’ont pas un grand besoin de cognition se contentent volontiers d’histoires divulgâchées[21].

L’échelle du besoin de cognition soulève toutefois une difficulté : c’est une autoévaluation, et les réponses socialement valorisées y sont assez évidentes. Quelqu’un qui veut faire bonne impression n’admettra probablement pas que « réfléchir, ça ne [l’]amuse pas du tout ». C’est pourquoi, au lieu de s’appuyer sur des autodescriptions, d’autres tests essaient de mesurer des compétences.

C’est le cas de l’échelle de compétence des adultes dans la prise de décision (*Adult Decision Making Competence scale*, ou A-DMC), qui mesure la tendance des individus à commettre des erreurs de jugement typiques, comme l’excès de confiance ou l’incohérence dans la perception du risque[22]. On peut citer également l’« Évaluation de la pensée critique de Halpern » (*Halpern Critical Thinking Assessment*), qui mesure la capacité de pensée critique, c’est-à-dire à la fois une disposition à la pensée rationnelle et un ensemble de compétences acquises[23]. Si vous vous soumettiez à cette évaluation, on vous poserait des questions comme celle-

ci : « Imaginez qu'un ami vous demande conseil pour choisir entre deux régimes proposant de perdre du poids. Un des régimes promet que les clients perdent en moyenne douze kilos, l'autre qu'ils en perdent quinze. Quelles sont les questions que vous poseriez à propos de ces régimes ? » En réponse, vous pourriez par exemple demander combien de gens ont perdu ce poids et s'ils n'en ont pas repris un an plus tard, ce qui manifesterait une bonne capacité de pensée critique. Les gens bien notés sur ces deux échelles semblent capables de meilleurs jugements dans la vraie vie : les scores plus élevés sont associés à une moindre fréquence d'événements indésirables, allant des simples contrariétés (comme être victime d'une « panne de réveil » ou acheter des vêtements qu'on ne portera jamais) aux problèmes sérieux (comme une grossesse non désirée ou une suspension de permis de conduire).

On pourrait faire l'hypothèse que ce type de mesures de style cognitif permet de prédire, de manière générale, la qualité du jugement. Leur pertinence semble cependant varier selon la nature du problème. Par exemple, dans le cadre des études sur la prévision que nous présenterons plus en détail au chapitre 21, Uriel Haran, Ilana Ritov et Barbara Mellers ont tenté d'identifier les styles cognitifs des meilleurs prévisionnistes. Mais ils ont conclu que le score obtenu sur le « besoin de cognition » ne permettait pas de prévoir qui allait se donner plus de mal pour trouver des informations supplémentaires[24], et plus généralement que ce trait n'était pas statistiquement associé à une performance plus élevée dans les tâches de prévision.

Selon ces chercheurs, la seule mesure de style cognitif ou de personnalité qui permette de prédire la performance en matière de prévision est une autre échelle, celle de l'« ouverture d'esprit active[25] », développée par le professeur de psychologie Jonathan Baron. Être activement ouvert d'esprit, c'est rechercher de manière proactive des informations qui contredisent nos hypothèses actuelles. Ces informations peuvent être les

opinions de personnes qui sont en désaccord avec nous, ou des éléments factuels nouveaux à mettre en balance avec nos convictions anciennes. Les individus activement ouverts d'esprit sont d'accord avec des propositions comme celle-ci : « On reconnaît une personne intègre au fait qu'elle accepte d'être convaincue par un argument contraire. » Ils ne sont pas d'accord, en revanche, avec l'idée que « changer d'avis est un signe de faiblesse » ou que « l'intuition est le meilleur guide pour prendre des décisions ».

Autrement dit, si les scores de réflexion cognitive et de besoin de cognition mesurent la propension à se livrer à la pensée lente et attentive, la pensée « activement ouverte d'esprit » va encore plus loin. Elle traduit l'humilité de ceux qui savent que leurs jugements sont des produits éternellement imparfaits, qui les remettent sans cesse sur le métier et qui souhaitent ardemment les corriger. Nous verrons que ce mode de pensée caractérise les prévisionnistes les plus chevronnés, qui révisent volontiers leurs convictions à la lumière des nouvelles informations disponibles. Détail encourageant, un certain nombre d'études suggèrent qu'être « activement ouvert d'esprit » est une compétence qui peut s'apprendre[26].

---

Notre propos n'est pas de tirer des conclusions définitives sur la manière de choisir des individus capables de produire de bons jugements dans un domaine donné. Deux idées se dégagent cependant de ce bref exposé. Premièrement, il serait sage que nous soyons plus systématiquement conscients de la différence entre les domaines où l'expertise peut être confirmée par comparaison avec des valeurs vraies (comme la prévision météorologique) et ceux où les seuls experts sont des experts d'estime. Un analyste politique peut être extrêmement éloquent et convaincant, tandis qu'un grand maître international d'échecs peut être timide et incapable d'expliquer le raisonnement qui lui inspire certains de ses coups. Nous devrions pourtant traiter le jugement professionnel de l'analyste politique avec plus de scepticisme que celui du joueur d'échecs : le premier est un

expert d’estime, tandis que les jugements du second sont constamment mis à l’épreuve des faits.

Deuxièmement, à formation et à expérience égales, certains juges seront meilleurs que leurs pairs, c’est-à-dire moins sujets à la fois au bruit et au biais. Parmi les nombreux facteurs qui expliquent ces différences, l’intelligence et le style cognitif ont une place importante. Et bien qu’il n’existe pas de mesure unique qui permette de prédire la qualité des jugements dans tous les domaines, on peut identifier un style cognitif associé aux bons jugements. La personne dotée de ce style cognitif cherche activement de nouvelles informations susceptibles de contredire ses convictions. Elle intègre méthodiquement ces nouvelles informations dans sa perspective. Et surtout, elle accepte, et souhaite même, que ces informations la conduisent à changer d’avis.

Il est intéressant de souligner que la personnalité que nous venons de décrire ne correspond pas au cliché du leader résolu. Le stéréotype du leader est, bien souvent, celui d’un chef qui décide vite et parle clair, tant il semble savoir, immédiatement et sans l’ombre d’un doute, quelle voie choisir. Ce genre de chef inspire confiance, bien sûr. Mais les données que nous avons passées en revue dessinent un autre profil. Quand l’objectif est de produire de bons jugements et de réduire les erreurs, mieux vaut que le chef (et tous ceux qui l’entourent) reste ouvert à des contre-arguments et soit pleinement conscient qu’il peut se tromper. S’il se montre résolu et déterminé, ce doit être au terme d’un processus de décision, pas au début.

## À propos des meilleurs juges

« Vous êtes un expert. Mais vos jugements sont-ils vérifiables, ou êtes-vous un expert d'estime ? »
« Nous devons choisir entre deux opinions, et nous ne savons rien de l'expertise de ces individus et de leurs résultats passés. Suivons le conseil du plus intelligent. »
« Mais l'intelligence n'est qu'une partie de l'histoire. La manière de penser est tout aussi importante. Peut-être devrions-nous choisir la personne la plus réfléchie et la plus ouverte d'esprit, pas seulement la plus intelligente. »

CHAPITRE 19

# Réduction des biais et hygiène de la décision

Un grand nombre de chercheurs et d'organisations se sont donné pour but de réduire les biais, ce qu'on appelle parfois « débiaiser » les jugements. Nous allons examiner ici quelques-uns de leurs principaux résultats[1], en distinguant les différents types d'interventions possibles, avant d'en examiner une qui nous semble mérite une attention plus approfondie. Nous nous tournerons ensuite vers la réduction du bruit et la discipline qu'elle appelle : l'hygiène de la décision.

## *Réduire les biais*, ex post *ou* ex ante

Un bon moyen de caractériser les deux principales approches de réduction des biais est de revenir à l'analogie de la mesure. Supposons que vous sachiez que votre pèse-personne ajoute, en moyenne, cinq cents grammes à votre poids chaque fois que vous vous pesez. Votre balance est manifestement biaisée, mais cela ne la rend pas pour autant inutile. Vous pouvez remédier au problème de deux manières : soit en ôtant cinq cents

grammes au chiffre que vous lisez à chaque pesée (mais c'est ennuyeux, et vous risquez d'oublier de le faire) ; soit en réglant le cadran une fois pour toutes pour rendre l'instrument plus exact.

Ces deux approches se retrouvent dans la réduction des biais de jugement. On peut réduire les biais soit *ex post*, en corrigeant les jugements après qu'ils ont été produits, soit *ex ante*, en intervenant avant le jugement ou la décision.

La correction des biais *ex post* est souvent pratiquée de façon spontanée et intuitive. Imaginons que vous supervisiez une équipe qui estime pouvoir mener à bien en trois mois le projet dont elle a la charge. Peut-être ajouterez-vous à cette estimation une marge de sécurité : en vous donnant, par exemple, quatre mois plutôt que trois, vous corrigerez le biais optimiste que vous supposez présent (le biais du planificateur).

On peut aussi pratiquer ce type de correction de manière plus formelle. Le ministère britannique des Finances a ainsi publié un guide pour l'évaluation des programmes et des projets[2]. Pour corriger les biais optimistes dans les prévisions de coût et de durée des projets, ce manuel suggère aux planificateurs d'appliquer des ajustements en pourcentage à leurs estimations initiales. En principe, ces pourcentages doivent refléter les biais d'optimisme constatés par le passé au sein de l'organisation concernée. S'il n'existe pas d'historique permettant de mesurer l'excès d'optimisme de l'organisation, le manuel fournit des pourcentages d'ajustement adaptés à chaque type de projet.

Quant à la prévention des biais *ex ante*, celle-ci se divise à son tour en deux grandes catégories d'approches. Certaines des plus prometteuses consistent à modifier « l'architecture du choix », c'est-à-dire l'environnement dans lequel le jugement ou la décision est produit. Ces modifications agissent comme des « coups de pouce » (*nudges*), qui vont soit réduire l'effet d'un biais, soit même exploiter un biais connu afin de produire une meilleure décision[3].

Un exemple de *nudge* largement utilisé aux États-Unis et au Royaume-Uni est l'inscription automatique dans les régimes de retraite facultatifs. Par inertie, par procrastination ou par biais d'optimisme, beaucoup de salariés ne font pas la démarche de s'inscrire dans ces plans d'épargne-retraite, et n'ont pas conscience des conséquences de cette inaction sur le niveau de vie qu'ils auront une fois retraités. L'inscription automatique fait en sorte que, par défaut, les employés épargnent pour leur retraite – tout en préservant leur liberté de ne pas le faire : il suffit pour cela qu'ils le choisissent délibérément. Ce *nudge*, qui s'est avéré très efficace pour augmenter les taux de participation, s'accompagne quelquefois de plans *Save More Tomorrow* (« Demain, épargner plus ») : les employés décident alors par avance d'affecter à leur épargne-retraite une partie de leurs futures augmentations de salaire. Là encore, la décision reste réversible, mais dans leur grande majorité, les salariés s'en tiennent à leurs choix initiaux. Le principe de l'inscription automatique (et plus généralement du choix par défaut) a fait ses preuves dans de nombreux autres domaines, par exemple pour encourager le choix d'utiliser de l'énergie verte, pour faire bénéficier de repas gratuits dans les cantines scolaires les enfants qui y sont éligibles, ou pour réduire le non-recours à divers programmes d'aide sociale.

Les *nudges* peuvent agir sur d'autres aspects de l'architecture du choix. Ils peuvent rendre la décision plus facile, par exemple en simplifiant les formalités administratives préalables à une prise en charge psychiatrique. Ils peuvent aussi mettre en relief certaines caractéristiques d'un produit ou d'une activité, par exemple en présentant de façon claire et explicite des frais auparavant dissimulés. Que ce soit dans une cafétéria, dans un supermarché, sur un formulaire administratif ou sur un site de vente en ligne, on peut agencer les options de manière à donner aux visiteurs un « coup de pouce » qui les aide à surmonter leurs biais. Si les aliments les plus sains, par exemple, sont présentés en bonne place et de façon bien

visible dans une cafétéria, il est probable que les clients seront plus nombreux à porter sur eux leur choix.

Bien distincte des *nudges*, l'autre approche de la réduction des biais *ex ante* consiste à former les individus pour qu'ils soient capables d'identifier et de surmonter leurs biais. Certains auteurs parlent de *boosting*[4], car ces interventions visent à développer les capacités des individus, par exemple en leur donnant des rudiments de statistique.

Former les gens à surmonter leurs biais est une entreprise louable, mais plus difficile qu'il n'y paraît. Bien sûr, il est toujours utile de recevoir une formation[5]. Ainsi, les personnes qui ont suivi pendant plusieurs années des cours de statistiques sont – heureusement – moins nombreuses à commettre des erreurs communes de raisonnement statistique. Mais il n'est pas pour autant facile d'apprendre à éviter les biais. Des dizaines d'années de recherche ont montré que même des professionnels solidement formés dans leur spécialité peuvent avoir du mal à transférer à d'autres domaines ce qu'ils y ont appris. En météorologie, par exemple, les prévisionnistes sont formés à calibrer leurs prévisions, et donc à éviter le biais d'excès de confiance. Quand de bons météorologistes annoncent 70 % de chances de pluie, il pleut, *grosso modo*, 70 % du temps. Mais si on pose à ces mêmes météorologistes des questions de culture générale, ils seront, comme tout un chacun, sujets à l'excès de confiance[6]. Apprendre à éviter un biais suppose d'abord de reconnaître une analogie entre le problème nouveau auquel on fait face et ceux qu'on a déjà rencontrés ailleurs, ensuite d'en conclure que le biais qu'on a constaté là-bas risque d'être à nouveau présent ici, et enfin de déployer, en les adaptant à la nouvelle situation, les stratégies qui ont permis de surmonter ce biais. Ce n'est pas facile.

Des chercheurs et des éducateurs sont parvenus à faciliter ce processus de transfert au moyen de techniques d'enseignement non traditionnelles. Une étude dirigée par Carey Morewedge, de l'Université de Boston, a par exemple utilisé une combinaison de vidéos éducatives et de « jeux

sérieux ». Les participants y apprenaient à reconnaître des erreurs causées par le biais de confirmation, l'ancrage et d'autres biais psychologiques. Après chaque jeu, ils recevaient un feed-back sur les erreurs qu'ils avaient commises et apprenaient des techniques permettant de les éviter. L'étude conclut que les jeux (et, dans une moindre mesure, les vidéos) réduisent le nombre d'erreurs commises par les participants, non seulement dans un test réalisé immédiatement après l'apprentissage, mais encore quand les participants répondent à des questions similaires huit semaines plus tard[7]. Une autre étude dirigée par Anne-Laure Sellier, d'HEC Paris, a montré que l'apprentissage d'un biais peut être transféré d'un domaine à l'autre. Des semaines après avoir appris à détecter le biais de confirmation dans un jeu vidéo éducatif, des étudiants en MBA faisaient face, dans un autre cours et sur un tout autre sujet, à un cas d'entreprise les exposant à ce même biais. L'étude suggère que les connaissances acquises dans le jeu vidéo les ont aidés à éviter ce piège[8], alors même que personne ne leur avait dit qu'il y avait un lien entre les deux exercices.

## *Une limite à la réduction des biais*

Qu'elles corrigent les biais *ex post* ou qu'elles les préviennent *ex ante* par le *nudge* ou le *boosting*, la plupart des approches de réduction des biais ont un point commun : elles visent un biais en particulier, qu'elles supposent présent. Cette hypothèse est raisonnable, mais pas toujours vérifiée.

Prenons à nouveau l'exemple de la planification de projet. On sait que le biais d'excès de confiance affecte les équipes projets de façon générale, et on peut donc raisonnablement supposer, quand on considère une équipe en particulier, que ce biais est présent. Mais on ne peut pas être certain que ce soit le seul biais qui affecte cette équipe, ni même le principal. Peut-être le chef d'équipe a-t-il récemment conduit un projet du même type, dans

lequel il a connu un échec cuisant dont le souvenir est encore vif. L'heuristique de disponibilité pourrait alors le conduire à être exagérément prudent et à se tromper dans son estimation – mais dans la direction opposée à celle que vous pensiez devoir corriger. Ou peut-être l'équipe a-t-elle réalisé ses prévisions par analogie avec un projet similaire et s'est-elle ancrée sur le temps qu'il avait fallu pour réaliser celui-ci. Autre possibilité : anticipant que vous vous donneriez une marge de sécurité supplémentaire, peut-être l'équipe a-t-elle prévenu votre ajustement en annonçant une estimation encore plus hardie que ce qu'elle pense réellement. Bien malin, alors, qui peut prévoir l'effet net de ces différents biais sur l'erreur de l'équipe.

La même complexité se retrouve dans d'autres types de décisions. Dans une décision d'investissement, par exemple, l'excès de confiance peut certainement jouer, et les exemples ne manquent pas. Mais un autre biais puissant, l'*aversion à la perte*, produit l'effet contraire : les décideurs hésitent souvent à prendre des risques, par crainte de perdre leur mise initiale. Il n'est pas certain *a priori* que le premier biais produise toujours un effet plus puissant que le second.

Autre cas, celui d'une entreprise qui prépare son budget et doit allouer des ressources à différentes unités. Les décideurs peuvent être à la fois trop optimistes quant aux perspectives des nouveaux projets que proposent certaines unités (l'excès de confiance, encore) et trop timides quand il s'agit de réduire les ressources allouées aux unités existantes (un problème provoqué par le *biais du statu quo*, qui, comme son nom l'indique, est notre préférence pour laisser les choses en l'état).

Tous ces exemples l'illustrent : il est bien difficile de savoir précisément quels biais psychologiques affectent un jugement donné et quel sera leur effet net sur l'erreur. Dès qu'un jugement est un tant soit peu complexe, une multitude de biais psychologiques peuvent être à l'œuvre, tantôt se combinant pour ajouter de l'erreur dans le même sens, tantôt conduisant à

compenser une erreur par une autre. Leurs conséquences deviennent vite imprévisibles.

On peut en conclure que la réduction des biais, qu'il s'agisse de correction *ex post* ou de prévention *ex ante* d'un biais spécifique, est utile, mais dans certaines situations seulement : celles où l'on sait quel biais corriger. C'est le cas si la direction générale de l'erreur est connue et se manifeste sous la forme d'un biais statistique clair. L'exemple type est l'erreur de prévision, résultat suffisamment robuste pour justifier qu'on corrige le biais du planificateur.

Le problème est considérablement plus délicat dans les très nombreuses situations où la direction probable de l'erreur n'est pas connue *a priori*. En particulier, dès lors que l'effet des biais psychologiques varie d'un juge à l'autre, il est, en pratique, imprévisible – et source de bruit systémique. Pour réduire l'erreur dans ce type de situation, corriger un seul biais à la fois n'est pas suffisant : si plusieurs biais sont à l'œuvre, il faudrait idéalement pouvoir les détecter tous.

## *L'observateur de la décision*

Nous voudrions suggérer ici que le bon moment pour cette détection des biais n'est ni avant ni après la décision, mais en temps réel. Bien sûr, les individus qui formulent un jugement n'ont généralement pas conscience de leurs propres biais au moment où ceux-ci les égarent. Cette inconscience est elle-même un biais bien connu, le *point aveugle des biais*[9]. Mais il est souvent assez facile d'observer les biais d'autrui. D'où cette suggestion : un observateur convenablement formé devrait être capable de se rendre compte en temps réel qu'un ou plusieurs biais connus affectent une décision ou une recommandation[10].

Pour illustrer cette démarche, imaginons un groupe d'individus réunis pour produire des jugements complexes et importants : un gouvernement face à une épidémie ou à une autre crise grave ; des médecins réunis pour décider du traitement d'un patient souffrant de symptômes complexes ; un conseil d'administration confronté à un choix décisif de stratégie d'entreprise. Imaginons maintenant un *observateur de la décision*, qui observe ce groupe d'individus et se sert d'une check-list pour déterminer si des biais empêchent ce groupe de s'approcher du meilleur jugement possible.

Le rôle de l'observateur n'est pas facile. Il ne fait même aucun doute que, dans certaines organisations, il est intenable. Donner l'alerte sur la présence de biais n'est possible que si celles et ceux qui prennent la décision finale sont fermement décidés à combattre ces biais. Par conséquent, l'initiative de mettre en place ce processus d'observation de la décision appartient aux décideurs finaux ; et ceux-ci devront aussi apporter à l'observateur un soutien sans faille. Nous ne recommandons évidemment à personne de s'autoproclamer observateur de la décision. Ce serait une bien mauvaise méthode pour se faire des amis et faire progresser sa carrière.

Des expériences informelles suggèrent toutefois que cette approche peut donner des résultats. Il faut pour cela que les dirigeants d'une organisation ou les responsables d'une équipe soient vraiment engagés dans cet effort, mais aussi que les observateurs de la décision soient bien choisis, tant pour leurs compétences que pour n'être pas eux-mêmes sujets à des biais importants sur la décision qu'ils observent.

Au moins trois profils d'observateurs de la décision peuvent répondre à ce cahier des charges. Dans certaines organisations, le rôle peut être confié à la responsable qui supervise la décision. En plus de s'intéresser au *contenu* des propositions qui lui sont soumises, la dirigeante se penchera de près sur le *processus* par lequel ces propositions ont été élaborées et sur la dynamique de l'équipe qui les a produites. Elle pourra ainsi prendre

conscience des éventuels biais qui ont pu affecter le travail de l'équipe dont elle a la charge.

D'autres organisations désignent, au sein de chaque équipe de travail, un membre qui fait office de « chasseur de biais » ; ce gardien du processus de décision rappelle en temps réel aux membres de l'équipe les biais susceptibles de les égarer. L'inconvénient de cette approche est que l'observateur de la décision est placé dans la position inconfortable de l'avocat du diable au sein de sa propre équipe, et risque donc de voir son capital politique s'épuiser rapidement.

Enfin, d'autres organisations recourent à un facilitateur extérieur, qui a l'avantage de la neutralité (et les inconvénients correspondants en termes de connaissances internes et de coûts).

Pour être efficace, l'observateur de la décision a besoin d'un minimum de formation et d'outils. En particulier, il lui faudra impérativement s'équiper d'une check-list des biais à détecter. Les check-lists ont fait leurs preuves en matière d'amélioration de la décision dans des contextes d'enjeux élevés, en particulier pour éviter la répétition d'erreurs passées [11].

En voici un exemple. Aux États-Unis, les agences fédérales sont tenues, avant de mettre en place de nouvelles réglementations, de réaliser une étude d'impact en bonne et due forme. Qu'il s'agisse de normes sur la pollution de l'air ou de l'eau, sur les accidents du travail, sur la sécurité alimentaire, sur la santé publique, sur les émissions de gaz à effet de serre ou sur la sécurité des frontières, l'étude d'impact doit démontrer que la réglementation proposée est nécessaire et proportionnée. Le contenu attendu de l'étude d'impact, codifié dans un épais document technique de cinquante pages dont le nom ne fait pas rêver (la circulaire OMB A-4), vise clairement à corriger les biais. Les agences fédérales doivent détailler la raison pour laquelle la réglementation est nécessaire, démontrer qu'elles ont envisagé des alternatives plus ou moins contraignantes, évaluer les coûts et les bénéfices de ces différentes options, présenter les informations de façon

non biaisée, et utiliser des taux d'actualisation appropriés dans leurs calculs des impacts futurs. Sur le papier, donc, les intentions étaient excellentes, mais cette énumération montre à elle seule que la tâche peut être fastidieuse. Dans les faits, de nombreuses agences ne se conformaient pas aux exigences de cet épais document technique. Peut-être personne ne l'avait-il même ouvert. La solution à ce problème a consisté à élaborer une check-list simple, qui tient sur une page et demie. La check-list rend bien plus facile de s'assurer que les agences se conforment aux principales prescriptions en matière d'évaluation des impacts[12].

Un exemple de check-list des biais qu'une observatrice de la décision pourrait utiliser figure dans l'annexe B[13]. Cette check-list générique n'est qu'une illustration, destinée à servir de point de départ : tout observateur de la décision voudra certainement développer une check-list sur mesure, afin d'en améliorer la pertinence et d'en faciliter l'adoption par son organisation[14]. En développant sa check-list, il lui faudra rester suffisamment sélectif : une check-list ne vise pas à dresser un inventaire exhaustif de tous les biais imaginables, mais au contraire à privilégier les plus fréquents et les plus sérieux.

Nous pensons que l'observation de la décision, aidée de check-lists « sur mesure », peut contribuer à limiter les biais. Cependant, si nous avons pu constater que cette démarche produit des résultats encourageants dans des essais informels, nous ne connaissons pas d'études ayant mesuré de manière systématique et à grande échelle les effets de cette approche, ni détaillé les avantages et les inconvénients des différentes manières possibles de la déployer. Nous espérons encourager à la fois les praticiens et les chercheurs à expérimenter davantage le recours à des observateurs de la décision pour réduire les biais en temps réel.

## *Réduire le bruit : l'hygiène de la décision*

Le biais est une erreur qu'on sait détecter et même expliquer. Un biais donné agit sur le jugement dans un sens prédéterminé : c'est pourquoi un *nudge* peut le contrecarrer. Il est prévisible : c'est pourquoi une action de formation peut aider à le combattre. Il est aussi souvent visible : c'est pourquoi un observateur peut espérer le diagnostiquer en temps réel.

Le bruit, en revanche, est une erreur dont l'occurrence et l'effet sont imprévisibles, et qu'il est difficile de détecter et d'expliquer. C'est pourquoi nous le négligeons si souvent, même quand il provoque des dégâts sérieux. Pour cette raison, la réduction du bruit est à la lutte contre les biais ce que l'hygiène préventive est aux traitements médicaux curatifs : son objectif est de prévenir, pas de guérir. Plus précisément, la réduction du bruit consiste à prévenir des erreurs potentielles – et indéterminées – avant qu'elles ne soient commises.

Nous appelons cette démarche de réduction du bruit l'*hygiène de la décision*. Quand vous vous lavez les mains, vous ne savez pas précisément contre quel pathogène vous vous prémunissez : il vous suffit de savoir que le lavage des mains est un geste prophylactique efficace contre un large éventail d'infections bactériennes ou virales. De même, l'hygiène de la décision consiste à adopter des techniques qui réduisent le bruit, donc l'erreur, sans jamais savoir précisément à quelle erreur en particulier cette prophylaxie permet d'échapper.

L'analogie avec le lavage des mains nous rappelle également que les mesures d'hygiène peuvent sembler fastidieuses. Les avantages qu'on en retire restent invisibles, car si tout se passe bien, vous ne saurez jamais à quel problème le lavage des mains vous a permis d'échapper. Inversement, si des problèmes surgissent, il ne sera pas toujours possible de savoir par qui et quand une norme d'hygiène a été enfreinte. Pour toutes ces raisons, il est difficile de faire respecter parfaitement les protocoles de lavage des mains, même chez des professionnels de santé pleinement conscients de leur importance.

Tout comme le lavages des mains et d’autres formes de prévention, l’hygiène de la décision est indispensable, mais ingrate. Corriger un biais clairement identifié donne au moins le sentiment d’avoir obtenu un résultat tangible. Hélas, la réduction du bruit ne vous offrira jamais cette satisfaction. Statistiquement, elle permettra d’éviter de nombreuses erreurs, mais vous ne saurez jamais lesquelles. Le bruit est un ennemi invisible, et prévenir l’assaut d’un ennemi invisible ne vous offrira jamais qu’une victoire invisible.

---

La réduction des biais est une démarche utile, même si elle n’est pas facile, à condition de savoir, dans une situation donnée, quel biais combattre. Contrairement à ce que l’on imagine parfois, ce n’est pas toujours le cas, car des biais multiples peuvent se conjuguer pour produire des effets imprévisibles. Une démarche possible pour contourner ce problème est de désigner un observateur de la décision, chargé d’identifier en temps réel les biais psychologiques.

Quant à la réduction du bruit, elle ne peut être que préventive, puisque les erreurs dues au bruit sont par nature imprévisibles. Comme les mesures d’hygiène sanitaire, cette hygiène de la décision est ingrate, mais indispensable. Les chapitres suivants introduiront les stratégies d’hygiène de la décision utilisées dans la police scientifique, la prévision, la médecine et les ressources humaines. Dans le chapitre 25, nous montrerons comment combiner ces stratégies dans une approche intégrée de la réduction du bruit.

## À propos de la réduction des biais et de l'hygiène de la décision

« Savez-vous quel est le biais que vous combattez et dans quel sens il affecte le résultat ? Si ce n'est pas le cas, cela veut probablement dire qu'il y a plusieurs biais à l'œuvre, et il est difficile de prédire lequel l'emportera. »

« Face à ce biais que vous avez identifié, pouvez-vous mettre en place une correction ? Un *nudge* ? Ou une mesure de formation ? »

« Avant de commencer à discuter de cette décision, désignons un observateur de la décision. »

« Avant de nous lancer dans cette décision, demandons-nous quelles mesures d'hygiène de la décision mettre en place. »

CHAPITRE 20

# Police scientifique : le séquencement de l'information

Le 11 mars 2004, à Madrid, plusieurs bombes placées dans des trains de banlieue tuaient cent quatre-vingt-douze personnes et en blessaient plus de deux mille. Les policiers espagnols relevèrent une empreinte digitale sur un sac en plastique trouvé sur place, et la transmirent *via* Interpol aux forces de l'ordre du monde entier. Quelques jours plus tard, aux États-Unis, le laboratoire criminel du Federal Bureau of Investigation (FBI) identifiait l'empreinte comme appartenant à Brandon Mayfield, un Américain résidant dans l'Oregon.

Mayfield faisait un suspect plausible. Ancien officier de l'US Army, il avait épousé une Égyptienne et s'était converti à l'islam. Devenu avocat, il avait notamment assuré la défense d'hommes mis en cause (et par la suite condamnés) pour avoir tenté de se rendre en Afghanistan afin d'y rejoindre les talibans. Il était sur une liste de suspects surveillés par le FBI.

Mayfield fut placé sous surveillance, sa maison fouillée et truffée de micros, ses téléphones mis sur écoute. N'ayant recueilli de la sorte aucune information utile, le FBI l'arrêta pour l'interroger. Mais Mayfield ne fut jamais formellement mis en examen. Il n'avait pas quitté le pays depuis dix ans. Qui plus est, pendant sa garde à vue, les enquêteurs espagnols, qui

avaient déjà fait savoir au FBI qu'ils ne pensaient pas que l'empreinte digitale pût correspondre à celle de Mayfield, l'identifièrent comme étant celle d'un autre suspect.

Mayfield fut relâché deux semaines plus tard. Le FBI lui fit des excuses, lui versa une indemnité négociée de 2 millions de dollars, et ordonna une enquête approfondie sur les causes de l'erreur. Sa principale conclusion tient en deux phrases : « L'erreur n'était pas méthodologique ou technologique. C'était une erreur humaine[1]. »

Heureusement, de telles erreurs humaines sont très rares. Elles n'en sont pas moins instructives. Comment les meilleurs experts américains ont-ils pu croire que l'empreinte était celle d'un homme qui n'avait jamais approché de la scène de crime ? Pour le comprendre, il faut d'abord se pencher sur l'examen des empreintes digitales et sur le jugement professionnel qu'il requiert. Comme nous allons le voir, l'identification des empreintes digitales n'est pas, comme on pourrait le croire, une question de fait qui relève de la science exacte. C'est en réalité un jugement, soumis aux biais psychologiques des experts qui le produisent. Ces biais peuvent créer bien plus de bruit, et donc d'erreur, qu'on ne l'imagine généralement. Et comme nous allons également le voir, certains laboratoires de police scientifique déploient pour faire face à ce problème une stratégie d'hygiène de la décision qui peut s'appliquer partout : le contrôle rigoureux du flux d'informations utilisées pour produire des jugements.

## *Les empreintes digitales*

Les empreintes digitales sont les impressions laissées par la friction du bout de nos doigts sur les surfaces que nous touchons. Si l'on trouve dès l'Antiquité des exemples d'empreintes digitales apparemment utilisées comme marques d'identification, l'exploitation moderne de l'empreinte

digitale remonte à la fin du XIX^e^ siècle, quand Henry Faulds, un médecin écossais, publia le premier article scientifique proposant de l'utiliser comme technique d'identification.

Dans les décennies qui suivirent, l'empreinte digitale conquit une place de plus en plus importante dans l'arsenal des techniques d'identification criminelle, remplaçant peu à peu les mesures anthropométriques inventées par Alphonse Bertillon (le « bertillonnage »). En 1912, Bertillon lui-même codifia un système formel pour comparer les empreintes digitales. Avant lui, sir Francis Galton, que nous avons déjà rencontré sur la foire aux bestiaux où il découvrit la sagesse des foules, avait développé un système similaire en Angleterre. (Ces pères fondateurs n'inspirent pas la sympathie : Galton pensait que les empreintes seraient un outil utile pour classer les individus en fonction de leur race, et Bertillon, sans doute victime de ses préjugés antisémites, apporta un témoignage d'expert décisif, et totalement faux, lors des procès d'Alfred Dreyfus en 1894 et 1899.)

Les policiers ne tardèrent pas à trouver aux empreintes digitales une autre utilisation. En 1892, Juan Vucetich, un policier argentin, fut le premier à comparer une trace digitale laissée sur une scène de crime avec le pouce d'un suspect. Depuis lors, il est devenu courant de recueillir des *traces digitales* (ou *traces papillaires latentes*), laissées par leurs propriétaires sur les lieux d'un crime, et de les comparer avec des *empreintes digitales* appartenant à des individus connus et recueillies dans des conditions contrôlées. L'ensemble des procédés qui permettent ces relevés et ces comparaisons (*dactylotechnie* et *dactyloscopie*) jouent un rôle déterminant dans les enquêtes policières, en permettant l'identification de suspects à partir des traces qu'ils laissent derrière eux.

Si vous êtes familier des lecteurs électroniques d'empreintes digitales, utilisés notamment aux frontières de nombreux pays, vous pensez peut-être que la comparaison des empreintes est une tâche mécanique et facilement automatisable. Mais comparer une trace digitale recueillie sur une scène de

crime avec une empreinte est beaucoup plus délicat que comparer deux empreintes. Quand vous appuyez fermement votre doigt sur un lecteur d'empreintes digitales, vous produisez une image nette et standardisée. Les criminels n'ont pas cette obligeance : les traces qu'ils laissent sont généralement floues, partielles, abîmées ou déformées. Elles en chevauchent d'autres, provenant ou non du même individu. Elles sont souillées par toutes sortes d'impuretés. Par conséquent, la quantité et la qualité de l'information que contient une trace digitale n'ont rien à voir avec celles d'une empreinte recueillie dans un environnement contrôlé. Décider si une trace correspond ou non à l'empreinte d'un suspect requiert un jugement. C'est le travail des experts qui les examinent.

Quand ils examinent un relevé de trace digitale, les experts suivent un processus appelé ACE-V : analyse, comparaison, évaluation et vérification. Ils doivent d'abord analyser la trace digitale pour savoir si elle est exploitable. Si c'est le cas, ils la comparent à l'empreinte d'un suspect. La comparaison conduit à une évaluation, qui peut aboutir à une *identification* (c'est le même individu), à une *exclusion* (ce n'est pas le même individu) ou à une décision *non conclusive*. En cas d'identification, une quatrième étape est déclenchée : la vérification par un autre expert.

Pendant des décennies, la fiabilité de cette procédure n'a jamais été remise en cause. Policiers et juges connaissaient les limites des témoignages oculaires et même celles des aveux obtenus d'un suspect, mais la preuve dactyloscopique (par les empreintes digitales) était, jusqu'à l'avènement de l'analyse de l'ADN, la forme de preuve la plus crédible. Jusqu'en 2002, personne n'avait réussi à contester sa valeur probante devant un tribunal américain. À cette époque encore, le site Web du FBI était catégorique[2] : « Les empreintes digitales sont un moyen *infaillible* d'identification personnelle. » Dans les très rares cas où survenaient des erreurs, elles étaient imputées à l'incompétence ou à la fraude.

Si la preuve dactyloscopique resta si longtemps incontestée, c'est en partie en raison de la grande difficulté de la mettre en défaut. La « valeur vraie » d'une trace digitale, c'est-à-dire l'identité de la personne qui a réellement été présente sur les lieux du crime, n'est *a priori* pas connue. Et entre la parole de l'expert qui a identifié un suspect et les dénégations de celui-ci, on n'hésite généralement pas très longtemps. Il est donc rarissime qu'apparaisse un conflit entre le jugement de l'expert et une vérité indiscutable.

Mais nous avons vu que le fait de ne pas connaître la vérité n'a rien d'inhabituel dans les problèmes de jugement, et qu'il n'y a pas besoin de connaître une valeur vraie pour mesurer le bruit. Quelle quantité de bruit y a-t-il dans l'analyse dactyloscopique ? Ou, plus précisément, puisque les experts en empreintes digitales, contrairement aux juges ou aux souscripteurs d'assurances, ne produisent pas un chiffre mais un jugement catégorique, quelle est la fréquence de leurs désaccords, et quelle en est l'origine ?

## *Le bruit occasionnel dans l'analyse des empreintes digitales*

La première personne à poser sérieusement ces questions a été Itiel Dror, un chercheur en sciences cognitives à l'University College de Londres. Dror a entrepris de réaliser l'équivalent d'une série d'audits de bruit dans un domaine où l'on croyait que le bruit n'existait pas.

Il peut paraître étrange de voir un spécialiste des sciences cognitives – un psychologue – mettre en cause le jugement d'experts en empreintes digitales. L'image que des séries télévisées comme *Les Experts* nous donnent de cette spécialité s'apparente à celle d'une science « dure », dont

les praticiens portent des gants en latex et manient le microscope électronique. Mais, pour Dror, examiner des empreintes digitales devait à l'évidence comporter une part de jugement. Et, en spécialiste des sciences cognitives, il connaissait le corollaire de cette observation : dès lors qu'il y a jugement, il y a bruit.

Pour tester cette hypothèse, Dror s'est d'abord intéressé au bruit occasionnel : la variabilité entre les jugements produits par le *même* expert examinant les *mêmes* empreintes en deux occasions différentes. Comme l'écrit Dror, « si des experts ne sont pas fiables au sens où ils ne sont pas constants dans leurs propres jugements, alors il faut s'interroger sur le fondement de leurs jugements et de leur professionnalisme[3] ».

Les empreintes digitales sont un terrain d'investigation parfait pour un audit de bruit occasionnel, car, contrairement aux cas que peuvent rencontrer un juge ou un médecin, les traces et les empreintes digitales ne sont pas faciles à mémoriser. Bien sûr, il faut laisser aux experts le temps d'oublier leurs jugements passés. En outre, pour être crédible, l'expérience doit avoir lieu dans le cadre du travail habituel des experts, afin que ceux-ci ne sachent pas qu'on est en train de tester leurs compétences. Dror a donc dû obtenir le consentement d'experts courageux et ouverts d'esprit : certains ont accepté qu'on leur resoumette à leur insu des cas déjà traités, sur une période pouvant aller jusqu'à cinq ans. Si, dans ce contexte, leurs jugements diffèrent de leurs conclusions initiales, c'est qu'il y a du bruit occasionnel.

## *Le biais de confirmation criminalistique*

Dans deux de ses premières études, Dror a complété son protocole en y ajoutant un détail important. Au moment où ils voyaient les empreintes pour la seconde fois, certains experts étaient aussi exposés à des

informations supplémentaires sur l'affaire. Par exemple, ceux qui avaient, la première fois, identifié les traces digitales comme appartenant à un suspect apprenaient cette fois que « le suspect a un alibi » ou que « l'analyse de l'arme à feu suggère que ce n'est pas lui ». Quand, en revanche, les experts avaient conclu, lors de leur premier examen, que le suspect pouvait être innocenté ou que les empreintes ne permettaient pas de conclure, on les informait que « l'inspecteur pense que c'est lui qui a fait le coup », qu'un « témoin oculaire l'a identifié » ou qu'il « a avoué le crime ». Dror parle de cette expérience comme d'un test de la sensibilité au biais des experts, car ces informations contextuelles étaient conçues pour activer dans un sens donné un biais psychologique (le biais de confirmation).

Et en effet, les experts se sont révélés sujets à ce biais. Quand les mêmes experts revoient des empreintes qu'ils ont déjà examinées dans le passé, mais en disposant cette fois d'informations de nature à biaiser leurs jugements, leurs conclusions changent parfois. Dans la première étude[4], cinq experts avaient rendu un verdict d'identification : quatre d'entre eux le modifièrent quand on leur représenta le même cas, assorti d'informations contextuelles qui suggéraient une exclusion. Dans la seconde étude[5], six experts devaient réexaminer quatre paires d'empreintes. Les informations contextuelles se traduisirent par des changements dans quatre décisions sur vingt-quatre. Certes, la majorité des décisions restèrent donc inchangées, mais pour ce type de décision, observer un changement injustifié dans un cas sur six est un niveau de bruit important. Par ailleurs, ces résultats, qui portent sur de petits échantillons, ont été reproduits indépendamment depuis par d'autres chercheurs.

Comme on peut s'y attendre, les experts changent plus facilement d'avis quand la décision est difficile que quand elle est facile. Ils se laissent plus facilement influencer quand les informations fournies pour les biaiser sont fortes que quand elles sont faibles. Et ils sont plus prompts à remplacer une décision conclusive par une décision non conclusive qu'à faire le

chemin inverse. Il est néanmoins troublant, note Dror, que « des experts en empreintes digitales prennent des décisions sur la base du contexte plutôt que sur celle des informations réellement contenues dans l'empreinte[6] ».

Les informations contextuelles peuvent donc biaiser les conclusions des experts ; mais ce n'est pas tout. Elles modifient en réalité *ce que l'expert perçoit*, et non seulement l'interprétation qu'il en donne. Dans une autre étude, Dror et ses collègues ont montré que des experts placés dans un contexte biaisé ne voyaient littéralement pas la même chose que d'autres experts qui n'avaient pas été exposés aux mêmes informations[7]. Lorsque la trace digitale qu'ils examinent est accompagnée d'une empreinte cible (celle d'un suspect), les experts relèvent beaucoup moins de détails (appelés *minuties*) que lorsque la trace leur est soumise toute seule. Une étude postérieure a confirmé cette conclusion, ajoutant que « la raison pour laquelle cela se produit n'est pas évidente[8] ».

Dror a inventé une formule pour désigner l'impact des informations qui biaisent les experts de la police scientifique : *le biais de confirmation criminalistique*. Ce biais a été confirmé depuis dans d'autres domaines de la police technique et scientifique et de la médecine légale, comme l'analyse des traces de sang, des débris d'incendie criminel ou des ossements, ainsi que la pathologie médico-légale. Même l'analyse de l'ADN – considérée comme le nouvel étalon-or de la police scientifique – peut être influencée par le biais de confirmation, au moins quand les experts doivent évaluer des mélanges d'ADN complexes[9].

La vulnérabilité de la police scientifique et des experts médico-légaux au biais de confirmation n'est pas qu'un problème théorique. En effet, dans la pratique, il n'existe pas de mesures systématiques pour s'assurer que les experts ne soient pas exposés à des informations de nature à biaiser leurs jugements. Au contraire, les experts reçoivent souvent ce type d'informations dans les courriers accompagnant les preuves qui leur sont soumises[10]. Ils sont aussi souvent en lien direct avec les policiers, avec les

procureurs, ou avec d'autres experts qui peuvent leur faire part de leurs propres impressions.

Le biais de confirmation pose un autre problème. Un bon moyen d'éviter les erreurs, formalisé par la dernière étape de la procédure ACE-V, est de faire procéder à une vérification indépendante par un autre expert avant de confirmer une identification. Toutefois, en règle générale, seules les identifications sont soumises à cette vérification. Il y a là un risque élevé de biais de confirmation, puisque l'expert chargé de la vérification sait que la première conclusion était une identification[11]. On ne peut donc pas considérer qu'une vérification concordante donne le niveau de confiance dans la conclusion qu'apporterait normalement l'agrégation de jugements indépendants : en réalité, les vérifications ne sont pas indépendantes.

C'est bien le biais de confirmation, et même une cascade de biais de confirmation, qui semble à l'origine du fiasco impliquant Brandon Mayfield. Il s'est trouvé non seulement deux, mais même trois experts du FBI pour concourir à l'identification erronée. Comme on peut le lire dans l'enquête effectuée plus tard à ce sujet[12], le premier semble avoir été impressionné par « la puissance de la corrélation » suggérée par le système automatisé de reconnaissance d'empreintes digitales qui avait identifié Mayfield. Même si l'expert n'avait apparemment pas connaissance du profil de Mayfield, les résultats fournis par le système informatisé, « couplés avec la pression inhérente au fait de travailler sur une affaire extrêmement médiatisée », ont été suffisants pour produire le biais de confirmation initial. Une fois cette identification faite par le premier expert, poursuit le rapport, « les analyses suivantes étaient contaminées ». Comme ce premier expert était un responsable chevronné et tenu en haute estime par ses pairs, « il est devenu de plus en plus difficile pour les autres experts de l'agence de s'opposer à lui ». L'erreur initiale fut reproduite et amplifiée, ce qui aboutit à une quasi-certitude que Mayfield était coupable. Fait éloquent, même un expert indépendant très respecté, chargé par le tribunal

d’examiner les preuves pour le compte des avocats de Mayfield, se rallia à l’avis du FBI et confirma l’identification[13].

Le même phénomène peut se produire dans d’autres disciplines de la médecine légale et de la police scientifique. L’identification par les empreintes digitales est considérée comme la plus objective de ces disciplines. Si les experts en empreintes peuvent être biaisés, alors ceux qui travaillent dans d’autres spécialités peuvent l’être aussi. Qui plus est, les experts peuvent s’influencer les uns les autres : si un expert en balistique sait que le détenteur de l’arme a été identifié par ses empreintes, son jugement peut s’en trouver biaisé. Et si un odontologue médico-légal sait que l’analyse de l’ADN a identifié un suspect, il sera moins enclin à suggérer que les marques de dents ne lui correspondent pas. Ces exemples illustrent le risque des cascades de biais[14] : comme dans les dynamiques de groupe évoquées au chapitre 8, une erreur initiale provoquée par le biais de confirmation devient l’information qui biaise un deuxième expert, dont le jugement en influence à son tour un troisième, et ainsi de suite.

Ayant établi qu’on peut faire varier le jugement des experts en les soumettant à des informations génératrices de biais, Dror et ses collègues se mirent en quête d’autres preuves de bruit occasionnel. Ils en découvrirent vite. Même quand les experts ne sont *pas* influencés par des informations nouvelles, il leur arrive de changer d’avis sur des empreintes qu’ils ont déjà examinées dans le passé[15]. Comme on peut s’y attendre, ces changements « sans raison » sont plus rares que les changements suscités par des informations qui biaisent les experts ; mais ils se produisent tout de même. En 2012, une étude commandée par le FBI a reproduit ce résultat à plus grande échelle en demandant à soixante-douze experts de se prononcer de nouveau sur vingt-cinq cas qu’ils avaient déjà examinés sept mois auparavant[16]. Dans ce large échantillon d’experts très qualifiés, le bruit occasionnel n’était pas nul : une décision sur dix environ se trouva modifiée.

Il faut souligner que la plupart des changements constatés consistaient à considérer comme non conclusive une comparaison précédemment jugée utilisable, ou l'inverse. Surtout, aucun changement ne déboucha sur une identification erronée. D'une certaine façon, l'étude met donc en évidence la prudence des experts. Elle implique néanmoins une observation préoccupante : certaines identifications par les empreintes digitales, qui n'ont pas été remises en cause et qui ont abouti à des condamnations, auraient pu, si elles avaient été examinées en une autre occasion, être jugées non conclusives. Car, même quand ce sont les mêmes experts qui examinent les mêmes empreintes, même quand le contexte n'a pas été modifié pour les biaiser, et même quand les conditions d'examen sont aussi inchangées que possible, leurs jugements ne sont pas parfaitement constants.

## *Un peu de bruit, mais combien d'erreurs ?*

Cette inconstance soulève une question capitale : le risque d'erreur judiciaire. Dans les procès criminels, le témoignage des experts en empreintes est souvent décisif. Or la stabilité de leur jugement est une condition nécessaire de la justesse de leur témoignage. En termes techniques, on dit que la *fiabilité intraobservateur* est nécessaire à la *validité*. En termes moins techniques, on ne peut pas être sûr d'être d'accord avec la réalité si l'on n'est pas d'accord avec soi-même.

Combien d'erreurs judiciaires sont imputables aux failles de la police scientifique ou de l'expertise médico-légale ? Une étude de l'Innocence Project[17], une association à but non lucratif qui s'attache à obtenir l'annulation de condamnations injustifiées, estime, après avoir analysé trois cent cinquante cas de disculpations, que la mauvaise application de la science médico-légale était un facteur d'erreur dans 45 % des cas. Cette statistique est certes inquiétante, mais la question qui se pose aux juges et

aux jurés est tout autre. Pour savoir dans quelle mesure ils peuvent faire confiance à un expert qui vient témoigner à la barre, ils ont besoin de connaître la probabilité que cet expert commette une erreur dans le cas présent (et non la probabilité que l'erreur d'un expert ait concouru à une erreur passée).

Le plus solide début de réponse à cette question se trouve dans un rapport préparé par le President's Council of Advisors on Science and Technology (PCAST), un organe consultatif rattaché à la présidence des États-Unis, qui réunit d'éminents scientifiques et ingénieurs. Ce conseil a publié en 2016 une étude approfondie sur l'utilisation de la science médico-légale et de la police scientifique en matière de justice pénale[18]. Le rapport comprend, entre autres, un état des lieux complet quant à la validité de l'analyse des empreintes digitales, et en particulier au risque de « faux positifs », c'est-à-dire d'identifications erronées, comme celle de Mayfield.

Les données disponibles sont étonnamment minces, et comme le note le PCAST, il est « déplorable » que des recherches pour en produire n'aient pas été réalisées plus tôt. Les chiffres les plus crédibles viennent de la seule étude à grande échelle publiée à ce jour sur la fiabilité de l'identification par les empreintes digitales, effectuée par des scientifiques du FBI en 2011[19]. L'étude impliquait cent soixante-neuf experts, chacun comparant une centaine de traces digitales avec des empreintes de suspects dont on savait, dans ces conditions expérimentales contrôlées, si elles correspondaient ou non à la trace relevée. Le principal résultat de l'étude était le faible taux d'identifications erronées : le taux de faux positifs était d'environ un sur six cents.

Un taux d'erreur d'un sur six cents peut sembler faible, mais, comme le relève le rapport du PCAST, il est « *beaucoup plus élevé* que ce que peut supposer le grand public (et, par extension, la plupart des jurés), sur la base des assurances données depuis longtemps sur l'exactitude de l'analyse des empreintes digitales[20] ». En outre, cette étude ne contenait pas

d’informations contextuelles de nature à biaiser les experts, et les experts participant à l’étude, qui avaient plusieurs semaines pour rendre leurs réponses, savaient qu’ils prenaient part à un test. Ces conditions éloignées de la pratique quotidienne peuvent se traduire par une sous-estimation du taux réel d’erreur. Une étude postérieure, réalisée en Floride, suggère un taux de faux positifs beaucoup plus élevé[21]. À tout le moins, ces disparités dans les résultats publiés mettent en évidence l’urgence de conduire des recherches plus poussées sur la fiabilité de l’analyse des empreintes digitales et sur la manière de l’améliorer.

Il y a au moins une conclusion rassurante qui semble toutefois se dégager des études réalisées à ce jour : les experts, quand ils se trompent, ont tendance à se tromper dans le sens de la prudence. Leur exactitude n’est pas parfaite, mais ils sont conscients des enjeux de leurs jugements, et ils prennent en compte l’asymétrie des conséquences de leurs erreurs. Étant donné la réputation de crédibilité de la technique des empreintes digitales, identifier à tort un suspect aura souvent des conséquences tragiques. Les conséquences des autres types d’erreurs ne sont pas aussi graves. Les experts du FBI observent ainsi que « dans la plupart des cas, une exclusion a les mêmes implications opérationnelles qu’une décision non conclusive[22] ». En d’autres termes, dire que la trace digitale trouvée sur l’arme du crime est celle d’un suspect est suffisant pour le condamner ; mais qu’on dise qu’une trace n’est pas la sienne ou qu’elle est inexploitable ne change généralement pas la donne, puisque aucune de ces deux décisions n’est suffisante pour le disculper.

D’autres éléments confirment que les experts font preuve de beaucoup de prudence et réfléchissent à deux fois – voire plus – avant de rendre une décision d’identification. Dans l’étude du FBI, quand les traces et les empreintes correspondaient, seuls un tiers des experts concluaient (à raison, donc) à une identification : les experts, note l’étude, « opèrent dans une culture où les faux positifs sont considérés comme plus graves que les faux

négatifs[23] ». D'autres études confirment que les experts font moins d'identifications erronées (faux positifs) que d'exclusions erronées (faux négatifs[24]). Et s'ils sont vulnérables aux biais, ils ne le sont pas dans la même proportion dans les deux sens. Comme l'écrit Dror, « il est plus facile de biaiser des experts dans le sens d'une décision prudente de non-conclusion que dans le sens d'une conclusion définitive d'identification[25] ».

En somme, les experts en empreintes digitales sont formés à considérer l'identification erronée comme le péché mortel à éviter à tout prix. Il faut leur reconnaître le mérite d'agir conformément à ce principe. On ne peut qu'espérer que leur niveau de prudence rende les cas d'identifications erronées, comme ceux de Mayfield et d'une poignée d'autres affaires très médiatisées, extrêmement rares.

## *Écouter le bruit*

Observer qu'il existe du bruit dans les jugements des experts de la police scientifique et de la science médico-légale n'est pas une critique de ces experts. Ce n'est que la conséquence directe de l'observation que nous avons maintenant faite à maintes reprises : dès lors qu'il y a jugement, il y a bruit, et plus qu'on ne l'imagine. L'étude des empreintes digitales est un travail qui semble objectif, à tel point que, bien souvent, nous ne le considérons pas spontanément comme une forme de jugement. Et pourtant, on trouve dans ces jugements de l'inconstance, du désaccord, et, parfois, des erreurs. Pour rares que soient celles-ci, elles existent, et, comme le note le rapport du PCAST, il serait bon que les jurys qui se reposent sur le témoignage d'experts en empreintes digitales en soient informés.

Le premier pas vers la réduction du bruit est de reconnaître son existence. Ce n'est, hélas, pas toujours chose facile. Beaucoup des experts de la police scientifique accueillirent les audits de bruit d'Itiel Dror avec un

grand scepticisme. En particulier, l'idée qu'un expert puisse se laisser influencer par des informations sur l'affaire en cours suscita l'ire de nombreux spécialistes considérés comme des autorités dans leur domaine. Dans une réponse à l'étude de Dror, le président de la Fingerprint Society, une association professionnelle britannique, écrit ainsi que « tout technicien qui […] est influencé dans un sens ou dans un autre dans son processus de décision […] fait preuve d'une telle immaturité qu'il/elle devrait aller chercher du travail à Disneyland[26] ». Dans le même esprit, le directeur d'un grand laboratoire médico-légal observe que l'accès à des informations sur l'affaire examinée – précisément le type d'informations dont Dror a montré qu'elles sont susceptibles de biaiser l'expert – « fournit une satisfaction personnelle qui permet [aux experts] d'apprécier leur travail *sans altérer en rien leur jugement*[27] ». Même le FBI, dans son enquête interne sur l'affaire Mayfield, notait qu'il « est courant que les experts réalisent des vérifications alors qu'ils connaissent les conclusions des experts précédents, *et que ces conclusions n'influencent pourtant pas les leurs*[28] ». Toutes ces remarques constituent un déni pur et simple de l'existence du biais de confirmation.

Même quand ils sont conscients en principe de ce risque, les experts de la police scientifique peuvent être victimes du point aveugle des biais, c'est-à-dire de la tendance à voir le biais chez autrui, mais pas chez soi. Dans une enquête à laquelle participaient quatre cents experts dans vingt et un pays[29], 71 % d'entre eux conviennent que « les biais cognitifs sont un sujet de préoccupation dans les sciences médico-légales en général », mais 26 % seulement que « leurs propres jugements sont influencés par des biais cognitifs ». Autrement dit, environ la moitié de ces scientifiques pensent qu'il y a du bruit dans les jugements de leurs collègues… mais pas dans les leurs ! Le bruit peut être un problème invisible, même pour les gens dont le travail est de voir l'invisible.

## *Séquencer l'information*

Pourtant, grâce à la persévérance de Dror et de ses collègues, les pratiques évoluent. Un nombre croissant de laboratoires médico-légaux a commencé à prendre des mesures pour réduire le risque d'erreur dans leurs analyses. Le rapport du PCAST salue par exemple le travail réalisé par le laboratoire du FBI pour réviser ses procédures afin de minimiser le risque de biais de confirmation.

Les étapes méthodologiques nécessaires pour ce faire sont relativement simples, et illustrent une stratégie d'hygiène de la décision qui peut trouver à s'appliquer dans de nombreux domaines : *séquencer l'information pour limiter la formation d'intuitions prématurées*. Dans tout jugement, certaines informations sont pertinentes et d'autres ne le sont pas. Avoir accès à plus d'informations n'est pas forcément un avantage. C'est particulièrement vrai quand ces informations risquent de biaiser le jugement en favorisant la formation d'une intuition prématurée.

Dans cet esprit, les nouvelles procédures instituées par certains laboratoires de la police scientifique visent à protéger l'indépendance de jugement des experts en leur donnant l'information dont ils ont besoin, seulement celle dont ils ont besoin, et seulement quand ils en ont besoin. En d'autres termes, le laboratoire les tient autant que possible dans l'ignorance de l'affaire, et ne leur divulgue les informations que progressivement. L'approche que Dror et ses collègues ont codifiée pour formaliser cette idée s'appelle la *divulgation séquentielle linéaire*[30].

Une autre recommandation de Dror procède de la même stratégie d'hygiène de la décision : il suggère que les experts documentent leurs jugements à chacune des étapes, c'est-à-dire par exemple qu'ils rédigent un compte rendu d'analyse d'une trace digitale *avant* d'examiner des empreintes pour décider si elles lui correspondent ou non. Cette séquence d'étapes vise à limiter le risque que les experts ne voient que ce qu'ils

cherchent. De même, les experts doivent enregistrer leur jugement sur les éléments qu'ils ont analysés *avant* d'avoir accès à des informations complémentaires sur l'affaire, dont on a vu qu'elles peuvent les biaiser. S'ils changent d'avis après avoir été exposés à ces informations, ces changements, et la raison qui les justifie, doivent être eux aussi expliqués par écrit. Toutes ces procédures limitent le risque qu'une intuition prématurée ne biaise la totalité du processus d'analyse.

La divulgation séquentielle des informations permet donc de préserver l'indépendance de jugement de chaque expert, en le protégeant de l'influence d'informations parasites. Dans le même esprit, Dror y ajoute une autre recommandation, qui vise, elle, à assurer que les jugements de différents experts soient indépendants les uns des autres. Quand un second expert est chargé de vérifier une identification, suggère Dror, il ne devrait pas savoir qu'il s'agit d'une vérification, afin de ne pas être influencé par le jugement du premier expert. Faire en sorte que des jugements multiples soient réellement indépendants est un aspect essentiel de l'hygiène de la décision, que nous retrouverons à maintes reprises.

---

La présence de bruit dans la science médico-légale et la police scientifique est bien sûr inquiétante, parce qu'il peut littéralement s'agir (au moins dans certains pays) d'une question de vie ou de mort. Mais elle est aussi révélatrice. Le fait que nous ayons si longtemps cru à l'infaillibilité des empreintes digitales montre à quel point notre confiance dans le jugement humain peut parfois être excessive. Et les études qui ont dissipé ce mythe témoignent qu'un audit de bruit peut révéler une quantité de bruit tout à fait inattendue. Fort heureusement, le cas de la police scientifique démontre aussi qu'il est possible, au moins dans une certaine mesure, de s'attaquer à ces problèmes en mettant en œuvre des changements procéduraux relativement simples. Cette conclusion devrait mettre du baume au cœur de

tous ceux qui se soucient de la qualité des décisions de leurs organisations, et les encourager à emprunter un chemin similaire.

La principale stratégie d'hygiène de la décision décrite ici – le séquencement de l'information – peut être utilisée dans de nombreuses situations pour prévenir le bruit occasionnel. Comme nous l'avons noté, le bruit occasionnel a un grand nombre de déclencheurs, qui vont de notre humeur du moment à la température extérieure. Nous ne pourrons donc jamais les contrôler tous avec précision. Mais nous pouvons tenter de mettre nos jugements à l'abri des éléments perturbateurs les plus manifestes. Vous savez déjà, par exemple, que vos jugements peuvent être altérés par la colère, la peur ou d'autres émotions. Et vous avez peut-être observé qu'il est souvent bénéfique, quand on en a le loisir, de reconsidérer un jugement passé, à un moment où les facteurs déclencheurs de bruit occasionnel peuvent être différents.

Si le rôle perturbateur de l'émotion semble évident, celui de l'information l'est beaucoup moins. C'est pourtant l'enseignement essentiel qu'on peut retirer de ce chapitre : l'information, *y compris quand elle est exacte*, est aussi un facteur de bruit occasionnel qui altère notre jugement. Comme dans l'exemple des experts en empreintes digitales, dès que l'on a accès à des informations qui ne devraient pas influencer notre jugement, ou dès que l'on sait ce que les autres pensent, notre jugement n'est plus indépendant. Le biais de confirmation peut nous amener à nous former une impression générale prématurée, puis à confirmer ce préjugement en ignorant les données incompatibles avec lui. La leçon qu'on peut en tirer pourrait s'inspirer de deux titres de films d'Alfred Hitchcock : un bon décideur devrait toujours essayer de conserver « l'ombre d'un doute », pour ne pas être « l'homme qui en savait trop ».

## À propos du séquencement de l’information

« Dès lors qu’il y a jugement, il y a bruit – même quand on ne se rend même pas compte qu’il s’agit d’un jugement, comme dans l’examen des empreintes digitales. »

« Nous avons beaucoup d’informations sur l’affaire, mais ne disons pas tout aux experts avant qu’ils n’aient produit leur jugement : même exactes, des informations dont ils n’ont pas besoin peuvent les biaiser. »

« Vous avez demandé une seconde opinion, mais si la personne qui vous la donne connaissait la première, elle n’est pas indépendante. Et c’est encore plus vrai pour la troisième : il peut y avoir une cascade de biais. »

« Pour combattre le bruit, il faut d’abord admettre qu’il existe. »

CHAPITRE 21

# Prévision : l’agrégation et la sélection

Quel sera le taux de chômage dans trois mois ? Combien de voitures électriques vendra-t-on l’an prochain ? Quels seront les effets du changement climatique d’ici 2050 ? Combien de temps faudra-t-il pour terminer ce nouvel immeuble ? Quel sera le résultat net de cette entreprise dans trois ans ? Ce nouvel employé fera-t-il bien son travail ? Quel sera le coût de la nouvelle réglementation contre la pollution de l’air ? Qui va gagner les élections ?

Toutes ces questions sont des questions de prévision. Des réponses qu’on leur apporte peuvent dépendre des choix fondamentaux, tant en matière de politiques publiques que dans le secteur privé.

Les chercheurs qui s’intéressent à l’exactitude des prévisions sont familiers de la distinction entre le biais et le bruit. Il est bien connu que les prévisions, dans certains contextes, souffrent de biais. Les agences officielles font par exemple montre d’un optimisme excessif dans leurs prévisions budgétaires[1]. Elles ont tendance, en moyenne, à faire des projections trop élevées du niveau de croissance économique, et des prévisions trop basses du niveau de déficit. On peut bien sûr spéculer sur la

part respective des biais psychologiques et des considérations politiques dans cet optimisme, mais le résultat est le même.

Indépendamment de l'excès d'optimisme, les prévisionnistes sont victimes d'un autre biais, plus fréquent encore : l'excès de confiance dans leurs prévisions, également appelé *excès de précision*. Quand on leur demande de formuler leurs prévisions sous la forme non pas d'estimations mais d'intervalles de confiance, ils ont tendance à choisir des intervalles plus étroits qu'ils ne le devraient[2]. Aux États-Unis, par exemple, une étude trimestrielle demande à des directeurs financiers de grandes entreprises d'estimer la croissance de l'indice S&P 500 pour l'année à venir[3]. Les directeurs financiers doivent donner deux chiffres : un minimum, en dessous duquel ils pensent que le résultat annuel n'a qu'une chance sur dix de se situer, et un chiffre maximum, au-dessus duquel le chiffre exact n'a, selon eux, pas plus d'une chance sur dix de se trouver. Les deux chiffres sont donc les bornes d'un intervalle de confiance à 80 %, ce qui veut dire que, collectivement, les directeurs financiers devraient voir juste 80 % du temps. Hélas, seuls 36 % des intervalles qu'ils proposent sont assez larges pour que le vrai niveau du S&P 500, une fois connu, s'y trouve. Les directeurs financiers sont bien trop confiants dans l'exactitude de leurs prévisions : ils sont victimes de l'excès de précision.

Les prévisionnistes sont aussi sujets au bruit. Un ouvrage de référence, *Principles of Forecasting*, de J. Scott Armstrong, rappelle que « le manque de fiabilité est une source d'erreur du jugement prévisionnel[4] », y compris chez les experts. Le bruit occasionnel est fréquent : les prévisionnistes varient souvent dans leurs jugements. Et bien sûr, ils ne sont pas toujours d'accord entre eux, même quand ce sont des spécialistes. Quand on demande à des professeurs de droit des pronostics sur les décisions de la Cour suprême des États-Unis, on constate qu'il y a beaucoup de bruit dans leurs prédictions[5]. Si l'on demande à des spécialistes de prévoir les bénéfices annuels d'une réglementation sur la pollution de l'air, on trouve

une variabilité très importante (par exemple, des estimations qui vont de 3 à 9 milliards de dollars[6]). Et quand des économistes font des prévisions sur la croissance ou le taux de chômage, on y trouve aussi beaucoup de variabilité. Le bruit dans les prévisions est partout, et la recherche sur la prévision en fournit une quantité impressionnante d'exemples[7].

La recherche propose aussi des solutions pour réduire le bruit et le biais. Nous ne passerons pas en revue ici toutes les stratégies de réduction du bruit, mais nous en examinerons deux qui ont l'avantage de pouvoir être très largement appliquées. L'une d'elles est celle que nous avons déjà évoquée au chapitre 18 : *sélectionner* de meilleurs juges. L'autre est sans doute la technique d'hygiène de la décision la plus universellement applicable : *agréger* un grand nombre d'estimations indépendantes. Commençons donc par un tour d'horizon des méthodes d'agrégation.

## *Agréger les prévisions pour les améliorer*

Le moyen le plus simple d'agréger plusieurs prévisions est d'en calculer la moyenne[8]. Le résultat est mathématiquement garanti : prendre la moyenne de *n* jugements divise le bruit par la racine carrée de *n*. Si, par exemple, vous utilisez la moyenne de cent jugements, vous réduirez le bruit de 90 %, et si vous faites la moyenne de quatre cents jugements, vous le réduirez de 95 % – ce qui revient pratiquement à l'éliminer. Cette loi statistique est le moteur de l'effet « sagesse des foules », présenté au chapitre 7.

Prendre la moyenne ne fait en revanche rien pour réduire le biais : puisque le biais est la moyenne des erreurs, il sera toujours aussi présent dans la moyenne des jugements. L'effet du calcul de moyenne sur l'erreur globale (l'erreur quadratique moyenne) dépend donc des parts respectives du bruit et du biais dans cette erreur globale : si l'erreur consiste

principalement en bruit, elle sera fortement réduite ; si elle consiste principalement en biais, l'amélioration sera bien moindre. C'est pourquoi la sagesse des foules fonctionne mieux quand les jugements de la « foule » sont indépendants les uns des autres : ils sont alors moins susceptibles d'être entachés d'erreurs partagées, c'est-à-dire de biais.

Empiriquement, de nombreuses preuves suggèrent que faire la moyenne d'un grand nombre de prévisions augmente fortement leur exactitude ; c'est par exemple le cas du « consensus » des analystes financiers. Qu'on parle de prévisions de ventes, de météorologie ou de prévisions économiques, la moyenne non pondérée d'un groupe de prévisionnistes fait mieux que la plupart d'entre eux, et souvent mieux que n'importe laquelle des prévisions individuelles[9]. On peut également calculer la moyenne de prévisions obtenues par différentes méthodes : dans une analyse faite sur trente comparaisons empiriques dans divers domaines, les prévisions combinées ont réduit les erreurs de 12,5 % en moyenne[10].

La moyenne simple n'est pas le seul moyen d'agréger des prévisions. Prendre la moyenne d'un *groupe sélectionné*[11], composé d'un petit nombre de juges (autour de 5), choisis sur la base de la fiabilité de leurs jugements récents, peut être aussi efficace que la moyenne d'une foule. Un avantage annexe de cette approche est que des décideurs qui respectent l'expertise font plus volontiers confiance à une stratégie qui ne repose pas seulement sur l'agrégation, mais aussi sur la sélection.

Une autre méthode pour produire des prévisions agrégées consiste à créer un *marché de prédictions*, où des individus prennent des paris sur les événements qui vont se produire, et sont donc incités à faire des prévisions justes. En général, les marchés de prédictions fonctionnent bien, c'est-à-dire que si le prix de marché d'un événement reflète une probabilité de 70 % que l'événement se produise, il se produira effectivement sept fois sur dix environ[12]. Un certain nombre d'entreprises de différents secteurs ont

expérimenté l'utilisation des marchés de prédictions pour obtenir des prévisions qui agrègent de nombreux points de vue[13].

Il est également possible d'agréger les jugements par un processus plus itératif, comme la *méthode Delphi*[14]. Sous sa forme classique, elle prévoit une succession de votes ou d'estimations que les participants développent séparément et soumettent anonymement à un modérateur. À chaque tour, les participants donnent les raisons de leurs estimations et réagissent aux raisons données par les autres, toujours anonymement. Le processus encourage la convergence des estimations (dans certaines versions, il la force même, en exigeant que les nouveaux jugements ne s'écartent pas trop de ceux des tours précédents). La méthode tire donc parti à la fois des avantages de l'agrégation et de ceux de l'apprentissage social.

La méthode Delphi est quelque peu complexe à mettre en place[15], mais il en existe une version plus simple, *mini-Delphi*, qui peut être appliquée dans une réunion unique[16]. Également appelée *estimation-discussion-estimation*, elle demande aux participants de produire d'abord chacun de leur côté une estimation, ensuite de l'expliquer et de la justifier au groupe, et enfin de produire une nouvelle estimation individuelle après avoir eu connaissance des estimations et des explications des autres. Le jugement du groupe est la moyenne des estimations individuelles obtenues lors de ce second tour.

## *Mieux comprendre la prévision : le Good Judgment Project*

En 2011, trois chercheurs de premier plan – Philip Tetlock (que nous avons rencontré au chapitre 11 à propos de l'évaluation des prévisions à long terme des événements politiques), son épouse Barbara Mellers et Don

Moore – ont créé le Good Judgment Project (GJP), qui est sans doute le programme de recherche le plus innovant et le plus ambitieux sur la prévision en général. Le but du projet est d'améliorer notre compréhension de la prévision et, en particulier, des raisons pour lesquelles certains individus s'y montrent meilleurs que la moyenne.

Le GJP a commencé par recruter des dizaines de milliers de volontaires, qui ne sont ni des spécialistes ni des experts, mais des gens ordinaires, sans autre point commun que leur intérêt pour ce projet. Ils ont dû répondre à des centaines de questions comme celles-ci :

– *La Corée du Nord fera-t-elle exploser une bombe nucléaire avant la fin de l'année ?*
– *La Russie va-t-elle annexer officiellement un autre territoire ukrainien dans les trois prochains mois ?*
– *L'Inde ou le Brésil vont-ils devenir membres permanents du Conseil de sécurité des Nations unies dans les deux années qui viennent ?*
– *Un pays va-t-il se retirer de la zone euro l'année prochaine ?*

Comme le montrent ces exemples, le projet s'intéresse aux grands événements géopolitiques ou économiques ; mais on peut supposer que les compétences qui permettent à certains individus de se distinguer quand ils répondent à ces questions seraient utiles aussi face aux problèmes de prévision plus prosaïques qui se posent dans tous les domaines. Et c'est bien à cette question que Tetlock et ses collègues se sont attaqués : celle de savoir si certains individus étaient des prévisionnistes particulièrement performants, et pourquoi. Ils ont aussi cherché à comprendre si l'aptitude à faire des prévisions pouvait être enseignée ou, au moins, améliorée.

Avant d'en venir à leurs conclusions, il nous faut d'abord expliquer quelques aspects de la méthode adoptée par Tetlock et son équipe pour évaluer les prévisionnistes. Premièrement, ils ont demandé aux participants

de produire beaucoup de prévisions. Sur un petit nombre de prévisions, la chance peut être déterminante : si vous prédisez que votre équipe de football favorite va gagner son prochain match, et qu'elle le gagne effectivement, cela ne fait pas nécessairement de vous un bon prévisionniste. Peut-être pariez-vous *toujours* que votre équipe favorite va gagner, et peut-être ne gagne-t-elle qu'une fois sur deux – auquel cas votre capacité de prévision n'a rien d'impressionnant. Les chercheurs ont donc jugé les participants à leur performance moyenne sur un grand nombre de prévisions.

Deuxièmement, les chercheurs n'ont pas demandé aux participants de formuler des prévisions binaires, du type « ceci se produira » ou « ceci ne se produira pas ». Dans le langage courant, c'est bien cela que « prévoir » un événement veut dire : prendre position dans un sens ou dans l'autre. Mais, compte tenu de ce que nous avons appelé l'ignorance objective des événements futurs, une prévision binaire est trop grossière. Le GJP a donc demandé aux participants d'estimer la *probabilité* que chaque événement se produise.

L'intérêt de ces prévisions probabilistes est qu'elles permettent de formuler des jugements plus fins. Si, par exemple, vous pensiez, en 2016, qu'Hillary Clinton avait 70 % de chances de gagner l'élection présidentielle, cela ne signifie pas nécessairement que vous êtes un mauvais prévisionniste. Les choses qui ont une probabilité de 70 % de se produire doivent, par définition, se produire 70 % du temps – et donc, *ne pas* se produire 30 % du temps, ce qui n'est pas rien. Pour évaluer la performance d'un prévisionniste qui utilise des probabilités, il faut donc regarder, sur un grand nombre de prédictions, si ses estimations de probabilité recoupent la réalité. Supposons par exemple qu'une prévisionniste, Margaret, évalue cinq cents événements différents, et attribue à tous ces événements une probabilité de se produire de 60 %. Si trois cents d'entre eux se produisent effectivement, alors nous pourrons en conclure que la confiance de

Margaret était bien *calibrée*. Le bon calibrage est une des conditions de la bonne prévision.

Troisièmement, Tetlock et ses collègues poussèrent un peu plus loin la subtilité. Ils ne demandèrent pas seulement à leurs prévisionnistes de faire une estimation de la probabilité qu'un événement se produise dans un délai donné (par exemple d'ici douze mois) ; mais ils leur donnèrent aussi la possibilité de réviser continuellement leurs prévisions à la lumière d'informations nouvelles. Supposons par exemple que vous ayez estimé, en 2016 encore, que le Royaume-Uni avait 30 % de chances de quitter l'Union européenne avant la fin de 2019. Au fil des sondages de plus en plus favorables au « Leave », vous auriez probablement voulu revoir votre prévision à la hausse. Au lendemain du référendum, il n'était toujours pas certain que le Royaume-Uni sortirait de l'Union à la date prévue, mais cela paraissait sûrement beaucoup plus probable (c'est finalement en 2020 que le Brexit s'est produit).

Chaque fois qu'apparaissait une nouvelle information pertinente, les prévisionnistes du GJP pouvaient donc mettre à jour leurs prévisions ; et chacune de ces révisions était considérée comme une prévision nouvelle prise en compte dans l'évaluation de leur performance. Ce mécanisme incite les participants à se tenir informés des actualités et à réviser leurs prévisions aussi souvent que nécessaire. C'est ce genre d'actualisation continue que l'on attend des prévisionnistes professionnels, que ce soit dans une entreprise ou dans un gouvernement : eux aussi doivent réviser leurs prévisions quand apparaissent des informations nouvelles, quitte à être critiqués parce qu'ils changent d'avis. (Une réponse à ce genre de critique, parfois attribuée à John Maynard Keynes, est restée célèbre : « Quand les faits changent, je change d'avis. Et *vous*, qu'est-ce que vous faites ? »)

Quatrièmement, pour évaluer la performance des prévisionnistes, le GJP s'est servi d'un système de notation spécifique, inventé en 1950 par Glenn

W. Brier, un météorologiste américain. Le *score de Brier* mesure l'écart entre ce que les gens prévoient et ce qui s'est vraiment passé.

L'intérêt du score de Brier est qu'il permet de contourner un problème récurrent de la prévision probabiliste : la tendance des prévisionnistes à « se couvrir » en évitant de prendre des positions trop tranchées dans un sens ou dans l'autre, et donc à toujours rester proches, dans leurs prévisions, de 50 %.

Pour illustrer ce point, revenons à Margaret, la prévisionniste qui avait estimé que cinq cents événements avaient une probabilité de 60 % de se produire. Comme trois cents d'entre eux s'étaient effectivement produits, nous avions jugé ses prévisions bien « calibrées ». Mais ce résultat n'est pas si impressionnant que cela si Margaret est météorologiste et si elle habite une ville où il pleut, en moyenne annuelle, 60 % des jours de l'année (comme Belém, au Brésil, ou Juneau, dans l'Alaska). Comme elle se contente de prévoir tous les jours une probabilité de pluie de 60 %, ses prévisions sont certes bien calibrées, mais elles sont assez peu utiles en pratique. Margaret vous dit seulement, en substance, que vous habitez une ville où il pleut souvent, et que vous avez intérêt à ne jamais vous séparer de votre parapluie. Comparons-la à Nicole, une météorologiste « idéale », qui prévoit, elle, une probabilité de pluie de 100 % pour les trois cents jours où il pleut effectivement, et une probabilité de 0 % pour les deux cents jours où il ne pleut pas. Le calibrage de Nicole est aussi parfait que celui de Margaret : dans les deux cas, quand la prévisionniste prédit une probabilité de pluie de X %, la pluie tombe exactement X % du temps. Mais les prévisions de Nicole sont infiniment plus utiles : elle vous dit clairement quand prendre votre parapluie, tandis que Margaret se contente d'ouvrir, au sens figuré, le sien ! Le terme technique qui caractérise la prévision de Nicole est qu'elle fait preuve, en plus d'un excellent calibrage, d'une *résolution* parfaite.

Les scores de Brier récompensent à la fois le calibrage et la résolution. Pour avoir une bonne note, vous ne devez pas seulement avoir raison en moyenne (c'est-à-dire être bien calibré) : vous gagnez aussi à prendre des positions tranchées en différenciant le plus possible vos prévisions (c'est-à-dire à avoir une résolution élevée). Les scores de Brier utilisent un calcul d'erreur quadratique moyenne, et les notes les plus faibles sont les meilleures : le score parfait serait de 0.

Maintenant que nous savons comment ils ont fait leurs prévisions et comment celles-ci ont été notées, quels ont été les résultats des volontaires du GJP ? Le résultat central de l'étude est que, si l'immense majorité d'entre eux ont des résultats médiocres, 2 % se distinguent par leur excellence. Ce sont ceux que Tetlock appelle les superprévisionnistes. Ils se trompent aussi, bien sûr, mais beaucoup moins que les autres. Et contrairement aux experts dont Tetlock avait étudié les prévisions et dont nous parlions au chapitre 11, ils font beaucoup mieux qu'une prévision « au hasard ». Tetlock rapporte même la remarque d'un haut responsable de l'administration américaine, qui s'étonne que les superprévisionnistes fassent « mieux que la moyenne des analystes des agences de renseignement qui ont accès à des messages interceptés et à d'autres données secrètes[17] ». Cette comparaison mérite qu'on s'y arrête. Les analystes des agences de renseignement sont des spécialistes confirmés de la prévision, pas des amateurs. Ils ont aussi accès à des informations classifiées. Et pourtant, ils ne font pas aussi bien que les superprévisionnistes.

## *Le bêta perpétuel*

Qu'est-ce qui explique la performance de ces superprévisionnistes ? Conformément à notre thèse du chapitre 18, vous pourriez soupçonner

qu’ils sont très intelligents. Vous n’auriez pas tort, mais ce n’est pas l’essentiel. Aux tests d’aptitudes mentales générales, les superprévisionnistes obtiennent des scores supérieurs à la moyenne des volontaires du GJP (elle-même très supérieure à la moyenne nationale), mais la différence n’est pas très importante, et beaucoup de volontaires qui se distinguent par une intelligence exceptionnelle ne sont pas pour autant des superprévisionnistes. Outre l’intelligence générale, vous pourriez aussi imaginer que les superprévisionnistes sont très à l’aise avec les chiffres. Et c’est vrai. Mais leur véritable avantage n’est pas leur don pour les maths : c’est l’aisance avec laquelle ils réfléchissent de façon probabiliste et analytique.

Plus précisément, ce qui caractérise les superprévisionnistes, ce sont leur propension et leur aptitude à structurer et à décomposer les problèmes. Quand ils sont face à une grande question géopolitique (Un pays va-t-il quitter l’Union européenne ? Une guerre va-t-elle éclater ici ou là ? Un responsable politique risque-t-il de se faire assassiner ?), ils ne produisent pas tout de suite un jugement global. Ils commencent par décomposer la question en plusieurs éléments : « Que faudrait-il croire pour que la réponse soit oui ? pour que la réponse soit non ? » Loin de s’en remettre à une réponse instinctive ou à une intuition globale, ils se posent toute une série de questions subsidiaires auxquelles ils essaient de répondre séparément.

Les superprévisionnistes excellent aussi à adopter ce que nous avons appelé la vision externe : ils sont très attentifs aux taux de base. Souvenez-vous de l’exemple de Michael Gambardi. Pour prévoir les chances de succès de l’individu, avant de se pencher sur les détails de son profil, il est utile de connaître le taux de base, c’est-à-dire la probabilité qu’un P-DG moyen soit encore en poste d’ici deux ans. Les superprévisionnistes sont de grands amateurs de taux de base. Quand on leur demande la probabilité qu’un différend frontalier conduise l’année prochaine à un affrontement armé entre la Chine et le Vietnam, les superprévisionnistes ne se demandent

pas seulement (ou pas tout de suite) si ces deux pays sont, aujourd'hui, en bons termes ou à couteaux tirés. Ils ont sans doute un avis sur la question, mais ils savent que leur intuition n'est généralement pas un bon point de départ pour estimer une probabilité. Ils commencent donc plutôt par chercher un taux de base : par exemple, la proportion des différends frontaliers qui dégénèrent en affrontements armés. Si de telles escalades sont rares, les superprévisionnistes partiront de ce fait, et c'est seulement ensuite qu'ils s'intéresseront de manière approfondie aux relations actuelles entre la Chine et le Vietnam.

Autrement dit, ce qui distingue les superprévisionnistes n'est pas leur intelligence brute : c'est leur *manière* de s'en servir. Les compétences qu'ils mobilisent témoignent du style cognitif « activement ouvert d'esprit » que nous avons évoqué au chapitre 18. Rappelez-vous le test qui mesure ce style cognitif : il contient des propositions comme « Les gens devraient prendre en considération les faits contraires à ce qu'ils croient » et « Il est plus utile de prêter attention aux gens qui ne sont pas d'accord avec vous qu'à ceux qui sont d'accord ». À l'évidence, les personnes « activement ouvertes d'esprit » n'hésitent pas, quand des informations nouvelles se présentent, à réviser leurs jugements (sans pour autant surréagir).

Pour caractériser le style de pensée des superprévisionnistes, Tetlock utilise l'expression « bêta perpétuel ». C'est le terme qu'utilisent les informaticiens pour désigner un logiciel qui n'est pas conçu pour devenir une version définitive, mais qui est continuellement testé et amélioré. Tetlock conclut que « l'indice le plus puissant qu'un participant deviendra un superprévisionniste est le bêta perpétuel : sa propension à actualiser continuellement ses prévisions et à améliorer sans cesse sa performance. […] Ce qui les rend si bons est moins ce qu'ils sont que ce qu'ils font : un laborieux travail de recherche, une réflexion prudente, une autocritique constante, la collecte et la synthèse d'autres points de vue, le jugement granulaire, la mise à jour permanente[18] ». Les superprévisionnistes sont

habitués à un cycle de pensée particulier : « essayer, échouer, analyser, adapter, essayer encore[19] ».

## *Le bruit et le biais dans la prévision*

Une question légitime que soulèvent ces observations est de savoir s'il est possible de former les personnes chargées d'établir des prévisions – par exemple dans les agences de renseignement – pour en faire des superprévisionnistes. Phil Tetlock et ses collaborateurs du GJP se sont donc attelés à cette tâche. Les résultats qu'ils ont obtenus apportent un éclairage complémentaire sur ce qui rend les superprévisionnistes exceptionnels.

La plus importante des études de Tetlock et de son équipe a consisté à diviser aléatoirement des participants ordinaires du GJP en trois groupes, afin de tester dans des conditions rigoureusement contrôlées l'effet de différentes interventions sur la qualité de leurs prévisions. Ces interventions illustrent trois stratégies distinctes d'amélioration du jugement :

1. *La formation* : certains des participants ont suivi une formation au raisonnement statistique et probabiliste. Ils y apprenaient en particulier le rôle que jouent différents biais (notamment le biais de confirmation et l'excès de confiance), ainsi que l'importance de calculer la moyenne de multiples prédictions issues de plusieurs sources, et la nécessité de prendre en considération les taux de base pertinents en cherchant à quelles « classes de référence » appartient le problème considéré.

2. *Le travail en équipe*, qui permet l'agrégation de jugements multiples : certains participants ont été regroupés en équipes, au sein desquelles tous les participants avaient connaissance des prédictions de leurs équipiers et pouvaient en débattre. L'idée sous-jacente était bien sûr que le travail en

équipe améliore la justesse des prévisions en encourageant les prévisionnistes à intégrer des arguments opposés à leur point de vue et donc à rester activement ouverts d'esprit.

3. *La sélection* : tous les participants ont été notés sur la justesse de leurs prévisions, et au bout d'un an, les 2 % qui obtenaient les meilleurs scores et qui étaient donc désignés « superprévisionnistes » ont eu la possibilité de travailler ensemble dans des équipes d'élite l'année suivante.

Les trois interventions ont toutes fonctionné, au sens où elles ont amélioré les scores de Brier des sujets. La formation améliore les scores, le travail en équipe les améliore encore plus, et la sélection est ce qui marche le mieux. Ce résultat important confirme donc l'intérêt des stratégies de sélection et d'agrégation des jugements.

Mais ce n'est pas tout. En se plongeant dans la masse de données recueillies sur les effets de chaque intervention, Ville Satopää, l'un des chercheurs qui collaborent avec Tetlock et Mellers, s'est attaché à comprendre *comment* chaque intervention améliore les jugements des participants. Nous ne tenterons pas de décrire ici la technique statistique sophistiquée qu'il a développée pour répondre à cette question, mais on peut en résumer l'idée sous-jacente[20]. En principe, a-t-il raisonné, les sources de différence dans la performance des prévisionnistes sont au nombre de trois :

1. Certains prévisionnistes peuvent être meilleurs que les autres pour trouver et analyser dans l'environnement des données pertinentes. En d'autres termes, ils peuvent avoir plus d'information : on peut s'attendre à ce que les prévisionnistes qui ont plus d'information soient plus performants.

2. Certains prévisionnistes peuvent avoir une tendance générale à se tromper dans un certain sens par rapport à la valeur vraie d'une prévision. Si, sur des centaines de prévisions, vous surestimez ou sous-estimez systématiquement la probabilité que certains changements se produisent (par rapport au *statu quo*), on peut dire que vous êtes sujet à une forme de biais, soit dans le sens du changement, soit dans le sens de la stabilité. Moins de biais doit se traduire par moins d'erreurs et de meilleures prévisions.

3. Certains prévisionnistes peuvent être moins susceptibles que les autres aux erreurs aléatoires. En matière de prévision, comme dans tous les jugements, le bruit peut provenir de nombreux facteurs : par exemple, les prévisionnistes peuvent surréagir à un certain type d'information (c'est un exemple de ce que nous avons appelé le bruit de pattern), être sujets au bruit occasionnel, ou utiliser l'échelle des probabilités de manière variable. L'ampleur et la direction de ces erreurs sont imprévisibles, mais les prévisionnistes qui y sont moins sujets devraient être plus performants : qui dit moins de bruit dit meilleure prévision.

Satopää, Tetlock, Mellers et leur collègue Marat Salikhov ont appelé cette tripartition le modèle BIN (pour *bias, information, noise*) : on peut devenir meilleur dans ses prévisions parce qu'on a plus d'information, parce qu'on a moins de biais, ou parce qu'on a moins de bruit. En suivant ce raisonnement, ils ont entrepris de mesurer le rôle de chacun de ces trois éléments dans l'amélioration de la performance des prévisionnistes lors de chacune des trois interventions que nous avons décrites plus haut, c'est-à-dire quand ils sont formés, regroupés en équipes, ou sélectionnés.

Leur conclusion est étonnamment simple : si les trois interventions ont fonctionné, c'est principalement en réduisant le bruit. Comme l'écrivent les chercheurs, « chaque fois qu'une intervention améliorait la justesse des prévisions, elle le faisait principalement en supprimant des erreurs de

jugement aléatoires. Curieusement, c'est même vrai pour l'action de formation, dont l'intention première était pourtant de réduire les biais[21] ».

Ce dernier point mérite d'être développé. Comme la formation visait à réduire les biais, on s'attendrait naturellement à ce que son effet principal soit la réduction du biais ; et pourtant c'est une réduction du bruit qu'on observe. Ce paradoxe s'explique par une distinction que nous avons déjà faite. La formation proposée par le GJP aux prévisionnistes était conçue pour les aider à combattre leurs biais *psychologiques*. Or, comme vous le savez maintenant, l'effet des biais psychologiques n'est pas toujours un biais statistique. Lorsqu'ils affectent différents individus ou différents jugements de différentes façons, les biais psychologiques sont une source de bruit. À l'évidence, c'est ce qui se passe dans ce concours : comme les événements à prévoir sont très variés, selon le sujet, les mêmes biais peuvent amener un prévisionniste à surréagir ou à sous-réagir. Il y a donc peu de chances que nous observions un biais *statistique*, défini comme la tendance générale d'un prévisionniste à croire trop souvent que des changements vont se produire (ou ne pas se produire). Au contraire, les biais psychologiques créent du bruit ; et c'est pourquoi apprendre aux prévisionnistes à lutter contre ces biais améliore leurs prévisions en réduisant le bruit qui les affecte.

Le travail par équipe produit un effet comparable à celui de la formation en matière de réduction du bruit, mais il a un résultat supplémentaire : il améliore aussi significativement la capacité des équipes à extraire de l'information. Ce n'est pas étonnant, si l'on considère la valeur qu'apporte l'agrégation de jugements multiples : plusieurs cerveaux qui travaillent ensemble sont mieux à même qu'un seul de trouver des informations. Si Alice et Brian travaillent ensemble, et si Alice a repéré des signaux que Brian n'a pas vus, leur prévision commune sera meilleure. Il est intéressant de noter que les superprévisionnistes, quand ils travaillent en équipe, semblent capables d'éviter les dangers de la polarisation de groupe et des

cascades informationnelles (évoqués au chapitre 8). Au contraire, ils mettent en commun leurs données, leurs idées et leurs intuitions, et, avec l'ouverture d'esprit active qui les caractérise, tirent le meilleur parti de ces informations combinées. Satopää et ses collègues expliquent cet avantage en ces termes : « Le travail en équipe – contrairement à la formation – permet aux prévisionnistes d'exploiter l'information[22]. »

L'intervention qui a le plus fort effet global est la sélection. Comme dans le travail par équipe, une partie de l'amélioration provient d'un meilleur usage de l'information. Les superprévisionnistes sont meilleurs que les autres pour trouver les informations pertinentes – ce qui suggère qu'ils sont peut-être plus intelligents, plus motivés et plus rompus à ce type d'exercice que le participant moyen. Mais le principal effet de la sélection, encore une fois, est de réduire le bruit. Les superprévisionnistes sont moins sujets au bruit que les individus ordinaires ou que les équipes ayant suivi une formation. Satopää et ses collègues le notent non sans étonnement : « Les "superprévisionnistes" semblent devoir leur réussite moins à une lecture éclairée des actualités qu'à une discipline supérieure en matière de réduction de l'erreur de mesure[23]. » C'est cette discipline de réduction du bruit que les prévisionnistes ordinaires semblent incapables d'égaler.

## *La valeur de la diversité dans la sélection et l'agrégation*

La réussite des superprévisionnistes met en lumière l'intérêt des deux grandes stratégies de réduction du bruit sur lesquelles ce chapitre met l'accent : la *sélection* (les superprévisionnistes sont, comme leur nom l'indique, supérieurs) et l'*agrégation* (quand ils travaillent en équipe, ils sont encore meilleurs). Ces deux stratégies sont applicables à de nombreux

jugements et tiennent une place centrale dans l'arsenal des techniques d'hygiène de la décision. Chaque fois que c'est possible, on peut, comme dans le GJP, les associer. Qu'il s'agisse de prévisionnistes, de recruteurs, d'experts en investissement ou d'arbitres de football, on peut assembler des équipes de juges soigneusement sélectionnés : d'abord pour être performants dans ce qu'ils font, et ensuite pour se compléter les uns les autres.

Ce dernier point est important. Jusqu'à présent, nous avons observé qu'on peut améliorer la qualité des estimations en prenant la moyenne de multiples jugements indépendants, comme dans les expériences sur la sagesse des foules. Nous avons vu aussi – et ce n'est pas surprenant – que quand la « foule » est composée d'individus supérieurement fiables, les résultats obtenus en agrégeant leurs jugements seront meilleurs encore. Mais ce n'est pas tout : on peut obtenir un gain d'exactitude supplémentaire en combinant des jugements qui sont non seulement indépendants mais aussi *complémentaires*[24].

Pour illustrer cette idée, imaginez que vous avez identifié quatre personnes qui ont été les témoins oculaires d'un crime. Il est bien sûr essentiel de s'assurer qu'elles ne s'influencent pas les unes les autres avant de témoigner : vous veillerez donc à les entendre séparément. Mais il semble évident que si, par ailleurs, elles ont vu le crime sous des angles différents, la qualité des informations qu'elles apportent n'en sera que meilleure. La diversité de leurs points de vue les rend complémentaires.

Pour décrire plus quantitativement l'apport de la complémentarité, on peut utiliser une autre analogie. La constitution d'une équipe de professionnels qui vont produire ensemble des jugements est comparable à l'assemblage d'une batterie de tests prédictifs – par exemple, pour prévoir le succès futur de candidats dans une formation ou un emploi. L'instrument standard qu'on utilise pour construire une batterie de tests est la régression multiple (que nous avons décrite au chapitre 9). Comme nous l'avons vu, la

régression multiple sélectionne des prédicteurs en fonction de leur valeur prédictive : le test qui prédit le mieux le succès sera donc sélectionné en premier. Mais le deuxième test retenu ne sera pas nécessairement celui qui arrive en deuxième position pour sa capacité à prédire indépendamment le succès. Ce sera celui qui *ajoute* le plus de force prédictive au premier test, en apportant des informations qui sont valides, mais pas redondantes avec les premières. Supposons par exemple que vous ayez à votre disposition trois tests : deux tests d'aptitudes mentales, qui ont une corrélation respective de 0,50 et 0,45 avec le succès futur ; et un test de personnalité dont la corrélation avec la performance n'est que de 0,30 – mais qui n'est pas corrélé avec les tests d'aptitude, alors que ceux-ci sont fortement corrélés entre eux. La solution optimale est de choisir d'abord le premier test d'aptitude (celui dont la validité est la plus élevée), puis le test de personnalité : pris séparément, celui-ci serait moins valide que le second test d'aptitudes mentales, mais il apporte de nouvelles informations.

La constitution d'une équipe de juges peut être considérée de la même manière[25]. Pour commencer, vous choisirez bien sûr le juge le plus « valide » au sens statistique, c'est-à-dire celui qui produit les jugements les plus justes. Mais votre choix suivant ne se portera pas nécessairement sur l'individu qui arrive deuxième dans le classement par ordre de validité : vous pourriez préférer un individu dont les prévisions sont moins exactes, mais qui apporte des compétences nouvelles. Une équipe sélectionnée de cette façon sera meilleure, parce que la validité des jugements qui sont mis en commun s'accroît plus rapidement quand les jugements ne sont pas corrélés les uns avec les autres que quand ils sont redondants.

Paradoxalement, quand une équipe est composée de cette manière, le bruit de pattern qu'on y observera sera relativement élevé : puisque les juges ne seront que faiblement corrélés entre eux, leurs jugements individuels de chaque affaire seront fréquemment différents. Mais si l'on utilise la moyenne des jugements produits par ce groupe (et non le jugement

d’un seul de ses membres), le résultat s’en trouvera amélioré : la moyenne d’un groupe au sein duquel il y a beaucoup de bruit sera plus juste que ne le serait la moyenne d’un groupe plus homogène.

---

L’étude des prévisions, et en particulier l’analyse des résultats du Good Judgment Project, enseigne l’importance d’agréger des jugements multiples, formulés par des individus judicieusement sélectionnés pour leur compétence, pour leur style cognitif, et pour leur diversité. Mais chaque fois qu’on parle d’agrégation, il faut renouveler une mise en garde importante. Quelle que soit la diversité des juges, l’agrégation de leurs jugements ne peut réduire le bruit que si ceux-ci sont réellement indépendants. Comme l’a souligné le chapitre sur le bruit des groupes, la *délibération* de groupe ajoute souvent davantage d’erreurs (sous forme de biais) qu’elle n’en supprime (en réduisant le bruit).

Pour tirer parti de la complémentarité des expertises et de la diversité des points de vue, une organisation doit donc s’attendre à des désaccords, inévitables quand les membres d’une équipe diverse produisent des jugements indépendamment les uns des autres. Faciliter et agréger des jugements à la fois indépendants et divers est souvent la stratégie d’hygiène de la décision la plus simple et la plus largement applicable. Mais elle exige qu’une organisation soit capable de composer avec des désaccords importants dans les jugements : il lui faudra les susciter, les entendre, les respecter, les applaudir, et finalement les résoudre.

## À propos de la sélection et de l’agrégation

« Faisons la moyenne de quatre prévisions indépendantes : c’est la garantie que le bruit sera réduit de moitié. »

« Comment pouvons-nous, comme les superprévisionnistes, rester activement ouverts d’esprit, en mode “bêta perpétuel” ? »

« Avant de parler de cette prévision, quel est le taux de base auquel nous devons nous référer ? »

« Nous avons les meilleurs dans cette équipe, mais avons-nous assez de diversité d’opinions ? »

« Vous avez entendu les jugements indépendants d’experts de différentes spécialités. Forcément, ils sont en désaccord : c’est une bonne chose. Maintenant, vous pouvez vous demander comment agréger ces jugements pour prendre votre décision. »

CHAPITRE 22

# Médecine : les directives

Il y a quelques années, un médecin généraliste (appelons-le le Dr Jones) diagnostiqua une forte tension artérielle à l'un de nos amis (appelons-le Paul). Il lui prescrivit un diurétique, mais le traitement n'eut aucun effet ; la tension de Paul resta élevée. Quelques semaines plus tard, le Dr Jones proposa un second traitement : un inhibiteur calcique. Son effet fut tout aussi modeste.

Ces résultats déconcertèrent le médecin. Après trois mois de visites hebdomadaires, la tension de Paul avait un peu baissé mais restait encore trop élevée. Quelle serait la prochaine étape ? Paul était inquiet et le Dr Jones décontenancé, d'autant que son patient était un homme assez jeune et par ailleurs en bonne santé. Il envisagea d'essayer un troisième traitement.

C'est alors que Paul déménagea dans une autre ville, où il consulta un autre généraliste (nous l'appellerons le Dr Smith). Paul lui fit part de ses problèmes de tension. Le Dr Smith répondit : « Achetez un tensiomètre, rentrez chez vous et contrôlez votre tension à domicile. Je ne pense pas que vous soyez hypertendu. Vous souffrez sans doute du syndrome de la blouse blanche : votre tension monte quand vous êtes chez le médecin ! »

Paul acheta un tensiomètre, rentra chez lui et prit sa tension : elle était normale. Elle l'est toujours depuis (et d'ailleurs, depuis que le Dr Smith lui

a parlé du syndrome de la blouse blanche, elle est normale aussi dans le cabinet de son médecin).

L'une des tâches principales des médecins est de faire des diagnostics : de juger si un patient est malade et, si c'est le cas, d'identifier la maladie. Tout diagnostic nécessite un jugement. Dans de nombreux cas, ce jugement est routinier et largement automatique, grâce à des règles et procédures qui minimisent le bruit. En général, par exemple, un médecin n'éprouve aucune difficulté à dire si quelqu'un s'est déboîté une épaule ou cassé un orteil. Il en va de même de problèmes plus techniques. Il y a assez peu de bruit quand on quantifie la dégénérescence d'un tendon[1]. De même, quand des pathologistes procèdent à une cytoponction pour détecter des lésions mammaires, leurs évaluations sont relativement simples et peu sujettes au bruit[2].

Par ailleurs, certains diagnostics ne nécessitent aucun jugement. La médecine progresse souvent en supprimant un élément de jugement, et en le remplaçant par une mesure ou un calcul. Pour diagnostiquer une angine bactérienne, un médecin va commencer par un test de dépistage rapide (TDR) de l'angine bactérienne à streptocoque, en utilisant un échantillon prélevé dans la gorge du patient. Il reste du bruit avec ces TDR, mais sans eux, il y en aurait beaucoup plus[3]. De la même manière, si vous avez, à jeun, un taux de glycémie de 126 milligrammes par décilitre (ou plus) ou un HbA1c (une mesure moyenne du taux de glycémie au cours des trois mois précédents) d'au moins 6,5 %, vous serez considéré comme ayant du diabète[4]. Au début de l'épidémie de Covid-19, les diagnostics que faisaient les médecins étaient largement affaire de jugement et fonction de la combinaison de symptômes observés ; à mesure que les tests de dépistage se sont répandus, ils ont rendu le jugement superflu.

Beaucoup de gens savent que le diagnostic est un jugement, et qu'il est donc sujet au bruit et à l'erreur. C'est pourquoi il est courant de conseiller au patient d'obtenir une seconde opinion. Dans certains hôpitaux, c'est

même obligatoire[5]. Chaque fois que la seconde opinion diverge de la première, il y a du bruit – mais il reste encore à savoir lequel des deux médecins a vu juste. Et bien sûr, chaque fois que le patient ne sollicite pas de seconde opinion, il y a du bruit aussi, même si personne ne s'en rend compte. C'est ce qui arrivait à Paul avec son premier médecin, et comme beaucoup de patients, Paul a été étonné de constater à quel point la seconde opinion pouvait diverger de la première. Car la surprise n'est pas le fait qu'il y ait du bruit dans la profession médicale : c'est son ampleur.

Notre objectif dans ce chapitre est d'approfondir ce constat et de décrire certaines des approches de réduction du bruit utilisées par le corps médical. L'étude du diagnostic médical nous fournira en particulier l'occasion de nous intéresser à une stratégie d'hygiène de la décision souvent employée : l'élaboration de *directives*. Nous n'ignorons pas qu'il serait possible d'écrire un livre entier sur le bruit en médecine et sur les diverses mesures que prennent les hôpitaux et les personnels soignants pour y remédier. Le bruit en médecine est d'ailleurs très loin de se limiter aux jugements diagnostiques, auxquels nous nous intéressons ici. Les traitements aussi sont sujets au bruit, et une abondante littérature scientifique en a fait état. Par exemple, depuis plus de vingt ans, le Dartmouth Atlas Project recense « les variations manifestes dans la manière dont les ressources médicales sont distribuées et utilisées aux États-Unis[6] ». On trouve des conclusions similaires dans un grand nombre de pays[7]. Pour notre propos, cependant, un bref examen du bruit dans les jugements diagnostiques sera suffisant.

## *Du bruit dans les diagnostics*

La littérature sur le bruit en médecine est considérable. Pour partie, elle est empirique, c'est-à-dire qu'elle mesure la présence de bruit. Mais une autre partie de cette littérature, fort heureusement, est prescriptive : elle suggère

des stratégies de réduction du bruit variées, qui sont une mine d'idées à prendre en considération dans de nombreux autres champs.

Quand il y a du bruit, il est possible qu'un médecin ait totalement raison et un autre totalement tort. À l'évidence, la compétence des diagnosticiens est essentielle. Une étude sur le diagnostic de la pneumonie par des radiologues, par exemple, a fait apparaître une forte quantité de bruit, dont l'essentiel est dû à des disparités de compétences[8]. Plus spécifiquement, dans ce cas, « la variation des compétences peut expliquer 44 % de la variation des décisions de diagnostic », ce qui suggère que « les mesures qui améliorent la compétence ont de meilleurs résultats que les directives uniformes » en matière de diagnostic. Ici comme ailleurs, la formation et la sélection sont évidemment d'une importance cruciale dans la réduction de l'erreur (qu'elle consiste en bruit ou en biais)[9].

Dans certaines spécialités, comme la radiologie et la pathologie, les médecins sont parfaitement conscients de la présence de bruit. Pour les radiologues, par exemple, la variation en matière de diagnostic est, de leur propre aveu, leur « talon d'Achille[10] ». Nous soupçonnons d'ailleurs que si le bruit en radiologie et en pathologie fait l'objet d'une attention particulière, ce n'est pas parce qu'il y en a beaucoup plus dans ces deux spécialités que dans d'autres, mais simplement parce que le bruit y est plus aisément mesurable. Il est relativement facile de réaliser des tests simples et clairs du bruit (et quelquefois de l'erreur) en radiologie ; par exemple en soumettant des scans ou des radios à un second praticien, ou en demandant à la même personne de les réexaminer pour réévaluer un premier diagnostic.

En médecine, on parle souvent de la variabilité interindividuelle comme d'un problème de *concordance interjuges*, qu'on mesure par un indicateur statistique, le *coefficient kappa*[11]. Plus le kappa est élevé, moins il y a de bruit. Un coefficient kappa de 1 reflète un accord parfait ; un coefficient de 0 signifie qu'il y a aussi peu d'accord qu'entre des jugements faits au

hasard, c'est-à-dire autant d'accord entre juges qu'entre des singes lançant des fléchettes sur une liste de diagnostics possibles. Certaines études sur le diagnostic médical font état d'une concordance « faible » ou « médiocre », ce qui veut dire qu'il y a beaucoup de bruit ; ou souvent d'une concordance « correcte », ce qui indique encore la présence de bruit.

Par exemple, sur la question de savoir quelles sont les interactions médicamenteuses[12] les plus cliniquement significatives, les médecins généralistes, après avoir examiné une centaine d'interactions sélectionnées de manière aléatoire, ont fait montre d'un accord « médiocre ». Quand il s'agit de diagnostiquer les différents stades de la néphropathie, les néphrologues ne montrent qu'un accord « faible à modéré[13] » : leurs jugements sur l'interprétation des tests standards utilisés pour évaluer des patients souffrant de cette pathologie peuvent être très différents. Sur la question de savoir si une lésion mammaire était cancéreuse, une étude n'a trouvé qu'un accord « correct[14] » entre des pathologistes. Sur le diagnostic de lésions mammaires proliférantes, l'accord, à nouveau, n'était que « correct[15] ». C'était aussi le cas de médecins analysant des IRM pour évaluer le degré de sténose du canal rachidien[16].

Ces résultats méritent qu'on s'y arrête. Nous avons dit que dans certains domaines, le niveau de bruit en médecine était très faible. Mais dans d'autres qui sont assez techniques, les médecins sont loin d'être exempts de bruit. Qu'une maladie grave soit ou non diagnostiquée chez un patient peut ainsi dépendre d'une sorte de loterie déterminée par le choix du médecin qui l'examinera.

Nous recensons ici quelques autres résultats issus de la littérature scientifique sur ce problème, dans des domaines où la quantité de bruit paraît particulièrement importante. Nous ne décrivons pas ces résultats pour tirer des conclusions définitives sur l'état de la pratique médicale, qui ne cesse d'évoluer et de s'améliorer (quelquefois très rapidement), mais pour donner une idée générale de l'omniprésence du bruit.

1. Les maladies du cœur sont la première cause de mortalité aux États-Unis[17], chez les hommes comme chez les femmes. L'angiographie coronarienne, une des principales méthodes utilisées pour diagnostiquer une pathologie cardiaque, évalue le degré d'obstruction des artères du cœur. Dans les cas non aigus, quand le patient souffre de douleurs récurrentes à la poitrine, on prescrit généralement un traitement – comme le placement d'un stent – quand une artère au moins est obstruée à plus de 70 %. Toutefois, l'interprétation des angiographies pour déterminer si une artère est obstruée à 70 % est sujette à une certaine variabilité, ce qui peut conduire à des procédures superflues (ou à l'erreur inverse). Une des premières études sur le sujet a montré que dans 31 % des cas, les médecins évaluant les angiographies n'étaient pas d'accord sur le point de savoir si une artère était obstruée à plus de 70 %[18]. Bien que les cardiologues soient largement conscients de cette variabilité dans la lecture des angiographies, et malgré des efforts constants et des mesures correctives, le problème n'a pas encore été résolu.

2. L'endométriose est une pathologie liée à la présence de tissu endométrial à l'extérieur de l'utérus. Elle peut être douloureuse et provoquer la stérilité. Elle est souvent diagnostiquée grâce à la cœlioscopie, c'est-à-dire l'introduction d'une petite caméra dans le corps. Dans une étude, 108 chirurgiens gynécologues ont visionné les images des cœlioscopies de trois patientes – deux qui souffraient d'endométriose à différents degrés de gravité, et une troisième qui n'en souffrait pas. On leur demanda ensuite d'estimer le nombre et l'emplacement des lésions. Leurs réponses divergèrent de façon considérable, tant sur le nombre que sur l'emplacement[19].

3. La tuberculose est l'une des maladies mortelles les plus répandues sur la planète : en 2016, elle a touché plus de dix millions de personnes et en a tué presque deux millions. La radiographie pulmonaire est largement

utilisée pour la détecter, et permet d'examiner les poumons pour y repérer des lésions évocatrices de la maladie. La variabilité de l'interprétation des images radiographiques et donc du diagnostic de la tuberculose est bien connue depuis près de soixante-quinze ans. Malgré des améliorations au fil du temps, des études ont continué de faire apparaître une variabilité significative, avec une concordance interjuges « modérée » ou seulement « correcte[20] ». Il y a aussi une forte variabilité dans le diagnostic de la tuberculose entre les radiologues de différents pays[21].

4. Dans une étude demandant à des pathologistes d'analyser des lésions cutanées pour y chercher la présence d'un mélanome – la forme la plus dangereuse de cancer de la peau –, l'accord entre eux n'était que « modéré ». Les huit médecins mobilisés dans cette étude n'ont été unanimes, ou en désaccord une seule fois, que dans 62 % des cas[22]. Une autre étude, réalisée dans un centre d'oncologie, a montré que la justesse du diagnostic du mélanome n'était que de 64 %, ce qui veut dire que les médecins faisaient un mauvais diagnostic pour près d'une lésion sur trois[23]. Une troisième étude a révélé que des dermatologues de la New York University, en examinant des biopsies de la peau, échouaient à diagnostiquer les mélanomes un peu plus d'une fois sur trois (36 % du temps). Les auteurs de l'étude concluaient que « s'agissant d'une maladie potentiellement mortelle, l'incapacité clinique à diagnostiquer correctement des mélanomes a des implications graves pour la survie des patients[24] ».

5. Il y a de la variabilité dans les jugements des radiologues en matière d'interprétation des mammographies pour le diagnostic du cancer du sein. Une étude importante a montré que le taux de faux négatifs chez différents radiologues allait de 0 % (le radiologue ne ratait jamais un cancer) à plus de 50 % (le radiologue jugeait à tort la mammographie normale plus d'une fois sur deux). De même, le taux de faux positifs allait de moins de 1 à 64 % (ce dernier chiffre signifiant qu'un radiologue, près de deux fois sur trois,

voyait un cancer là où il n'y en avait pas[25]). Ces faux négatifs et faux positifs, de la part de différents radiologues, sont des erreurs, dans lesquelles il entre du biais et du bruit.

En plus de ces exemples de bruit interpersonnel, largement présents dans la recherche existante, on trouve aussi des données sur le bruit occasionnel. Les radiologues sont quelquefois d'avis différents quand ils analysent la même image à plusieurs reprises[26], ce qui veut dire qu'ils ne sont pas d'accord avec eux-mêmes (moins souvent cependant qu'ils ne sont en désaccord avec des pairs). De la même façon, vingt-deux médecins auxquels on avait demandé de réévaluer le degré d'obstruction des artères sur des angiographies n'étaient en accord avec leur propre diagnostic que dans 63 à 92 % des cas[27]. Quand les critères sont vagues et les jugements difficiles, la *répétabilité intrajuges*, ainsi qu'on l'appelle, peut être médiocre[28].

Dans ces études, on ne trouve pas d'explication claire de ce bruit occasionnel. Mais une autre étude, où il n'est pas question de diagnostic, identifie une source de bruit qui mérite de retenir l'attention des patients comme celle des médecins[29]. En un mot, les médecins sont beaucoup plus enclins à prescrire des dépistages du cancer tôt le matin que tard dans l'après-midi. Dans un grand échantillon, le taux de prescription de tests de dépistage du cancer du sein et du côlon était le plus élevé à 8 heures du matin : 63,7 % ; il baissait tout au long de la matinée pour arriver à 48,7 % à 11 heures ; il augmentait de nouveau pour atteindre 56,2 % à midi puis retombait à 47,8 % à 17 heures. En d'autres termes, les patients qui entraient dans la cible des programmes de dépistage du cancer avaient nettement moins de chances de se le voir prescrire quand ils consultaient plus tard dans la journée.

Comment peut-on expliquer de tels résultats ? Une réponse possible est que les médecins prennent inévitablement du retard dans leurs rendez-vous quand ils reçoivent des patients souffrant de pathologies complexes qui

nécessitent davantage de temps que les vingt minutes habituelles. Nous avons déjà noté le rôle du stress et de la fatigue comme facteurs de bruit occasionnel (voir chapitre 7), et ces éléments semblent à l'œuvre ici : soucieux de respecter leur planning, certains médecins semblent se dispenser d'aborder avec leurs patients les questions de santé préventives. Une autre illustration du rôle de la fatigue chez les cliniciens est le taux inférieur de lavage des mains à la fin des gardes à l'hôpital[30].

## *Moins de bruit en médecine : l'intérêt des directives*

Une analyse exhaustive de la présence, de l'ampleur et des causes du bruit dans le diagnostic médical serait une contribution importante à la médecine et plus généralement à l'étude du jugement[31]. Elle n'existe pas encore à notre connaissance, et nous espérons que les travaux pour y parvenir se poursuivront. À ce stade, cependant, nous disposons d'un certain nombre d'indices sur les sources du bruit et sur les démarches qui permettent de le réduire.

À une extrémité du spectre, le diagnostic de certaines pathologies est quasiment automatique et ne laisse guère de place au jugement. Dans d'autres cas, le diagnostic, sans être automatique, est simple : tous les médecins ont de fortes probabilités d'arriver à la même conclusion. Dans d'autres cas encore, un certain degré de spécialisation – chez des spécialistes du cancer des poumons, par exemple – sera suffisant pour garantir que le bruit, sans disparaître, soit minimum. À l'autre extrémité du spectre, on trouve les cas qui laissent beaucoup de place au jugement, parce que les critères pertinents pour le diagnostic sont si ouverts que le bruit est

substantiel et difficile à réduire. Nous verrons que ces cas sont très fréquents en psychiatrie.

Comment réduire le bruit en médecine ? Comme nous l'avons vu, la formation peut élever le niveau de compétence, et de meilleures compétences, à l'évidence, réduisent le bruit[32]. Les médecins sont également conscients de la valeur qu'apporte l'agrégation de multiples jugements (notamment par les secondes opinions et les réunions de concertation)[33]. La technologie et, en particulier, les algorithmes offrent des solutions particulièrement prometteuses dans de nombreux domaines où l'on utilise déjà l'intelligence artificielle pour réduire le bruit dans les jugements diagnostiques. Ces algorithmes sont par exemple utilisés pour détecter des métastases sur les ganglions lymphatiques chez les femmes souffrant de cancer du sein : il s'avère que les meilleurs algorithmes sont supérieurs aux meilleurs pathologistes[34] – et ne sont évidemment pas sujets au bruit. Des algorithmes d'apprentissage profond sont aussi utilisés, avec un succès remarquable, pour détecter les problèmes oculaires liés à des diabètes[35]. Et l'intelligence artificielle fait déjà aussi bien que les radiologues pour détecter un cancer du sein à partir d'une mammographie[36] ; de futurs progrès la rendront probablement meilleure encore.

Les médecins feront très probablement un recours croissant aux algorithmes à l'avenir, car, en réduisant à la fois le bruit et le biais, la technologie promet de sauver des vies (et probablement de réduire les coûts du système de santé). Nous voulons cependant mettre ici l'accent sur un outil qui ne consiste pas à éliminer le jugement humain, mais à l'encadrer : les directives. L'exemple de la médecine montre en effet comment l'usage de directives peut produire d'excellents résultats dans certaines applications, mais aussi des résultats plus mitigés dans d'autres.

L'exemple le plus connu d'une directive de diagnostic médical est sans doute le score d'Apgar, conçu en 1952 par l'anesthésiste-réanimatrice

Virginia Apgar. Savoir si un nouveau-né se porte bien était naguère une question de jugement clinique pour les médecins et les sages-femmes. Le score d'Apgar leur a donné une directive pour y répondre. Il mesure la couleur de la peau du bébé, son rythme cardiaque, ses réflexes, son tonus musculaire et sa respiration, d'où le rétroacronyme Apgar : *apparence*, *pouls*, *grimace* (réflexes), *activité* (tonus musculaire) et *respiration*. Dans le test d'Apgar, chaque critère est noté 0, 1 ou 2. Le score total possible est donc de 10, mais un score de 7 ou plus suffit à indiquer une bonne santé (tableau 3).

| Catégorie | Nombre de points attribués |
|---|---|
| **Apparence (couleur de la peau)** | 0 : Tout le corps est bleu ou pâle<br>1 : Bonne couleur du corps mais pieds ou mains bleues<br>2 : Complètement rose ou couleur normale |
| **Pouls (rythme cardiaque)** | 0 : Aucun rythme cardiaque<br>1 : < 100 battements par minute<br>2 : > 100 battements par minute |
| **Grimace (réflexes)** | 0 : Aucune réponse quand les voies respiratoires sont stimulées<br>1 : Grimace pendant la stimulation<br>2 : Grimace et toux ou reniflement pendant la stimulation |
| **Activité (tonus musculaire)** | 0 : Inerte<br>1 : Certaines flexions des bras et des jambes<br>2 : Mouvement actif |
| **Respiration** | 0 : Pas de respiration<br>1 : Petit cri (gémissement, grognement)<br>2 : Fort, bon cri |

**Tableau 3.** Le score d'Apgar[37].

Il est important d'observer que, à l'exception du rythme cardiaque, qui est chiffré, chacun de ces critères implique un jugement. Mais comme le jugement global est décomposé en éléments individuels, et comme chacun de ces éléments est relativement facile à évaluer, même les praticiens n'ayant qu'un faible niveau de formation auront moins tendance à diverger. C'est pourquoi le score d'Apgar produit peu de bruit[38].

Le score d'Apgar illustre parfaitement de quelle manière les directives réduisent le bruit. À la différence des règles et des algorithmes, elles n'éliminent pas la nécessité du jugement : la décision n'est pas un calcul pur et simple. Des désaccords restent possibles sur chacun des critères, et par conséquent sur la conclusion finale. Mais les directives parviennent à réduire le bruit parce qu'elles décomposent une décision complexe en plusieurs sous-jugements plus simples sur des critères préétablis.

Les avantages de cette approche sont clairs si l'on considère le problème comme un exemple de prédiction, et si on lui applique la logique des « modèles simples » (linéaires) évoqués au chapitre 9. Quand elle porte un jugement sur la santé d'un nouveau-né, une clinicienne utilise et combine plusieurs indices prédictifs. Ce jugement peut être affecté par du bruit occasionnel : selon le jour, sa fatigue ou son humeur, la clinicienne pourrait prêter plus d'attention à certains prédicteurs ou moins à d'autres. Le score d'Apgar réduit cette variabilité en focalisant son attention sur les cinq critères empiriquement reconnus comme les plus importants. Il donne ensuite une indication claire de la manière dont les cinq critères doivent être évalués, ce qui simplifie grandement le jugement de chacun d'entre eux, et réduit donc le bruit. Il indique enfin comment pondérer mécaniquement les prédicteurs afin de produire le jugement global requis, là où différents cliniciens humains donneraient des poids différents à chaque critère. Le choix des bons prédicteurs, la simplification du jugement sur chacun d'eux, et l'agrégation mécanique : tout cela diminue le bruit.

Une approche analogue a été utilisée dans de nombreux domaines de la médecine. Le score de Centor, qui guide le diagnostic de l'angine streptococcique, en est un bon exemple. Le score additionne un point pour chacun des symptômes ou signes suivants : absence de toux, présence d'exsudats amygdaliens (taches blanches au fond de la gorge), ganglions mous ou gonflés dans le cou, température supérieure à 38 °C. En fonction du nombre de points recueillis par le patient, un prélèvement dans la gorge pour diagnostiquer la pharyngite streptococcique pourra être recommandé. L'évaluation et la notation sont relativement simples si l'on utilise cette échelle[39], qui a effectivement réduit le nombre de personnes soumises à des tests inutiles ou recevant un traitement superflu. (En anglais, les critères de ce score forment, comme le score d'Apgar, un rétroacronyme des quatre premières lettres du nom de Robert Centor.)

De même, des directives ont été élaborées pour le diagnostic du cancer du sein avec le *Breast Imaging Reporting and Data System* (BI-RADS), qui réduit le bruit dans l'interprétation et la rédaction des comptes rendus de mammographies. Une étude a confirmé que le BI-RADS accroît le niveau de concordance interjuges dans la lecture des mammographies, montrant ainsi que les directives peuvent réduire efficacement le bruit dans un domaine où la variabilité est grande[40]. En pathologie aussi, beaucoup d'efforts ont été faits, et avec succès, pour développer des directives[41].

## *Le cas déprimant de la psychiatrie*

En matière de bruit, la psychiatrie est un cas extrême. Quand des psychiatres examinent le même patient avec les mêmes critères de diagnostic, il est fréquent qu'ils soient en désaccord. C'est pourquoi les psychiatres ont fait de la réduction du bruit une priorité dès les

années 1940[42]. Mais, comme nous le verrons, les directives n'ont apporté qu'une aide modeste en la matière, en dépit de constantes améliorations.

Dès 1964, une étude portant sur quatre-vingt-onze patients examinés par dix psychiatres expérimentés montrait que la probabilité d'un accord entre deux opinions était seulement de 57 %[43]. Une autre, portant sur quatre cent vingt-six patients d'hôpitaux publics aux États-Unis, examinés de façon indépendante par deux psychiatres, n'a montré d'accord sur la maladie mentale présente que dans 50 % des diagnostics. Une autre encore, impliquant cent cinquante-trois patients en consultation externe, ne trouva que 54 % d'accord. Ces études n'analysent pas les sources du bruit ; mais elles notent que certains psychiatres étaient plus enclins à formuler certains diagnostics. Ainsi, certains avaient plus souvent tendance à diagnostiquer une dépression, d'autres un trouble anxieux.

Comme nous allons le voir, les niveaux de bruit restent élevés en psychiatrie. Pour quelle raison ? Les spécialistes n'ont pas de réponse unique et claire (les explications du bruit sont donc elles-mêmes sujettes au bruit). Sans doute la très grande variété des diagnostics possibles est-elle un facteur.

Un autre indice est fourni par une étude où, pour répondre à cette question, des chercheurs ont demandé à un premier psychiatre de s'entretenir avec un patient, puis à un second de mener un autre entretien après une courte période de repos[44]. Les deux psychiatres se sont ensuite réunis et, s'ils étaient en désaccord, ont discuté pour essayer d'en comprendre les raisons.

Une raison fréquemment invoquée était la « variabilité du médecin » : différentes écoles de pensée, différentes formations, différentes expériences cliniques, différentes méthodes d'entretien. Le même épisode d'hallucinations, par exemple, pouvait être interprété par un « clinicien ayant reçu une formation développementale comme un symptôme post-traumatique faisant suite à des violences subies », un autre « ayant une

orientation biomédicale pouvait expliquer les mêmes hallucinations comme un élément du processus schizophrénique[45] ». Ces différences entre médecins dans leur réponse à des cas spécifiques sont des exemples caractéristiques de bruit de pattern.

Au-delà des différences entre médecins, cependant, la principale raison du bruit était « l'inadaptation de la nomenclature ». Ces observations, et le mécontentement général de la profession vis-à-vis de la nomenclature psychiatrique, ont conduit à la révision de 1980 (la troisième édition) du *Diagnostic and Statistical Manual of Mental Disorders*, plus connu sous l'acronyme abrégé de *DSM-III*. Pour la première fois, le manuel contenait des critères explicites et détaillés pour le diagnostic des troubles mentaux, première étape vers l'introduction de directives.

Le *DSM-III* permit une augmentation considérable du nombre d'études sur le bruit dans les diagnostics psychiatriques[46]. S'il s'avéra utile pour le réduire, il fut loin d'y parvenir complètement[47]. Même après le *DSM-IV* (originellement publié en 1994) et son importante révision en 2000, le niveau de bruit restait élevé[48]. Comme le note une étude d'Ahmed Aboraya et de ses collègues, « l'utilisation de critères de diagnostic pour les troubles psychiatriques accroît la fiabilité des diagnostics[49] » ; et pourtant le risque demeure élevé que « les admissions d'un même patient se traduisent par de nombreux diagnostics différents[50] ».

Une nouvelle version du manuel, le *DSM-5*, a été publiée en 2013[51]. L'American Psychiatric Association espérait qu'elle permettrait de réduire le bruit parce qu'elle reposerait sur des critères plus objectifs et mieux calibrés[52]. Pourtant, le bruit dans le jugement des psychiatres demeure élevé[53]. Samuel Lieblich et ses collègues estiment ainsi que « les psychiatres ont de grandes difficultés à s'entendre sur qui souffre ou non d'un trouble dépressif grave[54] ». Les essais de terrain du *DSM-5* ont fait apparaître « un accord minimal[55] », ce qui « signifie que dans les conditions de l'étude, des psychiatres spécialisés très bien formés n'ont été d'accord

pour dire qu'un patient avait une dépression qu'entre 4 % et 15 % du temps ». Selon certains essais de terrain, le *DSM-5* a même aggravé les choses, montrant un niveau de bruit accru « dans tous les grands domaines, certains diagnostics, comme le trouble associant l'angoisse et la dépression […], étant si peu fiables qu'ils étaient inexploitables en pratique clinique[56] ».

Il semble que la principale raison du succès limité des directives soit qu'en psychiatrie, « les signes diagnostiques de certains troubles sont toujours vagues et difficiles à rendre opérationnels[57] ». Les directives réduisent le bruit en décomposant le jugement en critères pour lesquels il y a moins de désaccord, mais dans la mesure où ces critères demeurent subjectifs et ouverts à l'interprétation, le bruit reste probable.

Pour y remédier, des psychiatres éminents ont proposé des directives plus standardisées. Celles-ci prévoient[58] : (1) de clarifier les composantes du diagnostic en évitant les critères insuffisamment précis ; (2) d'élaborer des « définitions de référence » des symptômes et de leur niveau de gravité, l'idée étant que « des cliniciens qui sont d'accord sur la présence ou l'absence de symptômes auront plus de chances de s'accorder sur le diagnostic » ; et (3) d'utiliser, outre la conversation libre avec les patients, des techniques d'entretien structuré. Un projet de guide d'entretien prévoit par exemple vingt-quatre questions de dépistage, afin de permettre un diagnostic plus fiable de l'anxiété, de la dépression et des troubles du comportement alimentaire, entre autres.

Ces mesures semblent prometteuses, mais il reste à savoir dans quelle mesure elles permettront de réduire le bruit. Comme le soulignait un observateur, « le fait de se reposer entièrement sur les symptômes subjectifs du patient, l'indispensable interprétation de ceux-ci par le clinicien, et l'absence de mesure objective (comme un test sanguin) portent en germe l'absence de fiabilité du diagnostic des troubles psychiatriques[59] ». En ce

sens, il est possible que la psychiatrie se révèle irréductiblement réfractaire à la réduction du bruit.

---

La plupart des diagnostics médicaux comprennent une part importante de jugement. C'est ce qui les expose au bruit : les désaccords entre praticiens sont fréquents, et l'inconstance chez les mêmes praticiens n'est pas rare (notamment sous l'effet de la fatigue). Le niveau de bruit diminue, et l'exactitude des diagnostics progresse, quand l'automatisation remplace le jugement.

Mais la médecine a aussi beaucoup progressé sans supprimer le jugement, quand elle a réussi à le codifier et à l'encadrer par des directives. Les directives ont montré une grande efficacité pour réduire le bruit, le biais, et donc l'erreur – même s'il semble plus difficile d'en développer dans certaines spécialités, notamment en psychiatrie. Les directives qui existent rendent de grands services aux professionnels de santé et aux patients, et contribuent grandement à améliorer la santé publique. La médecine gagnerait sans doute à en développer plus encore[60], et d'autres domaines à s'en inspirer.

## À propos des directives

« Chez les médecins, le niveau de bruit est beaucoup plus élevé que nous ne le pensions. Dans le diagnostic du cancer, dans celui des maladies cardio-vasculaires, ou même dans la lecture des radiographies, il arrive aux spécialistes de ne pas être d'accord. Cela signifie que le traitement proposé au patient peut être le résultat d'une sorte de loterie. »

« Les médecins (et leurs patients) aiment à croire qu'ils prennent la même décision que l'on soit lundi ou vendredi, tôt le matin ou tard l'après-midi. Or il s'avère que ce qu'ils disent et font dépend souvent de leur niveau de fatigue. »

« Partout où il est possible de développer des directives, elles peuvent permettre aux médecins de faire moins d'erreurs. C'est bon pour les patients, bien sûr, comme pour le corps médical. Et c'est une technique d'hygiène de la décision dont d'autres domaines pourraient utilement s'inspirer. »

CHAPITRE 23

# Évaluation de la performance : le choix de l’échelle

Commençons par faire un petit exercice. Prenez trois personnes de votre connaissance – des amis ou des collègues, par exemple. Notez-les sur trois critères : la bienveillance, l’intelligence et l’ardeur au travail. Utilisez pour ce faire une échelle allant de 1 à 5 (1 étant la plus mauvaise note et 5 la meilleure). Demandez ensuite à quelqu’un d’autre qui connaît ces trois personnes – votre conjoint, votre meilleur ami, un proche collègue – de le faire à son tour.

Il y a de très fortes chances que certaines de vos notes soient différentes. Si vous et votre comparse le souhaitez, vous pouvez en discuter pour essayer d’en trouver les raisons. Peut-être la réponse réside-t-elle dans la manière dont vous avez utilisé l’échelle – une manifestation de ce que nous avons appelé le bruit de niveau. Ce serait le cas si l’un de vous deux pense qu’un 5 sur le critère « bienveillance » doit être réservé aux candidats à la béatification, tandis que l’autre distribue les 5 sans barguigner. Mais peut-être s’agit-il aussi de bruit de pattern, si vous différez dans vos perceptions respectives des degrés de bienveillance de ces trois personnes, voire dans vos définitions de cette qualité : parle-t-on de bonté, de générosité, de

dévouement ; ou simplement de gentillesse quotidienne, d'amabilité et d'égalité d'humeur ?

Imaginez maintenant que vous et votre comparse participiez à des évaluations de performance dans une entreprise qui valorise la bienveillance, l'intelligence et l'ardeur au travail, et que, pour les trois personnes que vous notez, une promotion ou une prime soit en jeu. La différence entre vos évaluations serait-elle aussi grande que dans le premier exercice ? Ou plus grande encore ? En tout état de cause, s'il y a des différences dans le niveau et le pattern de vos réponses, ce sera une source de bruit. Et c'est ce que l'on observe partout dans l'évaluation de la performance en entreprise.

## *Un travail de jugement*

Dans la quasi-totalité des grandes organisations, la performance est évaluée de façon formelle et régulière. Le moins qu'on puisse dire est que les personnes évaluées n'apprécient pas cette expérience. Comme le résumait un grand quotidien américain, « les études montrent que pratiquement tout le monde déteste les revues de performance[1] ». Pratiquement tout le monde a aussi conscience que, dans une certaine mesure au moins, ces évaluations sont sujettes à la fois au biais et au bruit. Mais la plupart des gens ne réalisent pas dans quelle mesure. Car, nous allons le voir, ce bruit est assourdissant.

Dans un monde idéal, on pourrait évaluer la performance des individus avec des données objectives. Il n'y aurait pas besoin de jugement. Mais les organisations modernes n'ont pas grand-chose à voir avec la fabrique d'épingles d'Adam Smith, où l'output de chaque ouvrier était parfaitement mesurable. Quel serait cet output pour une directrice financière ou une responsable de la recherche ? Dans un monde de « travailleurs de la

connaissance », les objectifs sont multiples, difficilement objectivables, et parfois contradictoires entre eux. Se focaliser sur un seul objectif peut produire des évaluations erronées et créer des incitations perverses. Le nombre de patients reçus chaque jour par un médecin est un facteur important de la productivité d'un hôpital, par exemple, mais il ne serait pas souhaitable que les médecins se focalisent sur ce seul indicateur, et encore moins qu'ils soient évalués et rémunérés sur cette seule base. Même quand il existe des indicateurs chiffrés de performance, comme les ventes d'un commercial ou le nombre de lignes de code écrites par un programmeur, ceux-ci doivent être replacés dans leur contexte : toutes les ventes ne sont pas aussi faciles à conclure, et tous les projets de logiciel ne sont pas identiques. Pour toutes ces raisons, un grand nombre de gens ne peuvent pas être évalués exclusivement sur des critères de performance objectifs. D'où l'omniprésence des évaluations de la performance basées sur le jugement[2].

## *Un quart de signal, trois quarts de bruit*

Des milliers d'articles de recherche ont été publiés sur la pratique des évaluations de performance. La plupart des chercheurs concluent que ces évaluations sont excessivement sujettes au bruit[3]. Cette inquiétante observation provient pour une bonne part de l'analyse d'évaluations de la performance dites « à 360 degrés ». Dans une évaluation à 360 degrés, la même personne est évaluée, généralement sur plusieurs critères de performance, par plusieurs évaluateurs : un ou plusieurs supérieurs, des pairs hiérarchiques, des subordonnés, voire des partenaires extérieurs (comme des clients ou des fournisseurs). On peut ainsi comparer plusieurs évaluations de la même personne sur la même période.

Les résultats ne sont pas encourageants. La part de la variance totale des évaluations qui peut être attribuée à la performance de la personne – la variance « vraie » – ne représente pas plus de 20 à 30 % de la variance totale. Tout le reste, soit 70 à 80 % de la variance des évaluations, est du bruit systémique[4].

D'où provient ce bruit ? Grâce aux nombreuses études décomposant la variance des évaluations de performance, nous savons que toutes les composantes du bruit systémique sont présentes[5]. Il est assez facile de se les représenter en prenant l'exemple de deux évaluatrices, Lynn et Mary.

Tout d'abord, si Lynn est indulgente et Mary sévère, c'est-à-dire que Lynn, sur l'ensemble des personnes évaluées, donne des notes en moyenne plus élevées que Mary, il y a du bruit de niveau. Comme nous l'avons noté à propos des juges judiciaires, ce bruit peut signifier soit que Lynn et Mary diffèrent réellement dans les impressions qu'elles forment, soit qu'elles n'utilisent pas de la même façon l'échelle d'évaluation pour exprimer des impressions identiques. Que la différence soit dans le niveau de leurs impressions ou dans celui de leurs expressions, le bruit de niveau résultant sera le même.

Ensuite, si Lynn vous évalue et qu'il se trouve qu'elle a une piètre opinion de vous et de votre travail, son indulgence habituelle pourrait s'effacer devant cette réaction négative à votre égard. C'est ce que nous avons appelé un pattern stable : la réaction personnelle d'un évaluateur donné à un individu donné. Ce pattern est propre à Lynn et à son interaction avec vous : si vous aviez été évalué par Mary, qui note tout le monde sévèrement mais n'a rien contre vous en particulier, le résultat aurait été bien différent. Il y a donc du bruit de pattern stable.

Enfin, supposons que Mary a découvert, juste avant de remplir le formulaire d'évaluation, que sa voiture a été emboutie sur le parking de l'entreprise. Au même moment, Lynn apprenait sa propre évaluation de performance, et se voyait récompensée d'une prime aussi inattendue que

généreuse. Les exemples sont extrêmes, bien sûr, mais ils illustrent le fait que l'humeur des évaluateurs, comme mille autres facteurs, peut produire du bruit occasionnel.

Comment le bruit systémique se décompose-t-il entre ces trois éléments (bruit de niveau, bruit de pattern stable et bruit occasionnel) ? Différentes études arrivent à des conclusions différentes, et il est très plausible que cette décomposition varie grandement d'une organisation à une autre. Toutes les formes de bruit n'en sont pas moins indésirables. Le message principal qui se dégage de ces travaux est simple : la plupart des « évaluations de performance » ont avec la performance de la personne évaluée un rapport beaucoup plus ténu qu'on ne l'imagine. Comme le résument les auteurs d'une synthèse de ces études, « la relation entre la performance au travail et l'évaluation de la performance au travail a de fortes chances d'être faible ou, au mieux, incertaine[6] ».

L'une des explications parfois avancées est que les évaluateurs ne communiquent pas dans leurs évaluations leur vraie perception de la performance des employés[7]. Il existe en effet de nombreuses raisons de soupçonner que les évaluateurs ne jugent pas toujours leurs employés objectivement, mais les notent « stratégiquement[8] », en fonction de leurs propres objectifs. Les évaluateurs ne manquent pas de mobiles : ils peuvent vouloir donner une meilleure note pour s'épargner une discussion pénible, pour favoriser quelqu'un qui attend depuis longtemps une promotion, voire, paradoxalement, pour se débarrasser d'un maillon faible qui ne peut être muté dans un autre service qu'en faisant état d'une bonne évaluation. Bref, les évaluateurs peuvent tricher, mentir, et faire de la politique.

Il ne fait aucun doute que ces calculs stratégiques affectent les évaluations ; mais ils ne sont pas, loin s'en faut, la seule explication de leur variabilité. Il y a du bruit dans le jugement des évaluateurs, même quand celui-ci est parfaitement sincère. La preuve en est fournie par une sorte d'expérience naturelle : celle des systèmes de feed-back à 360 degrés qui

sont utilisés à titre exclusivement développemental, c'est-à-dire que les notes (et le feed-back qui les accompagne) ne sont pas utilisées à des fins d'évaluation. Dans ce contexte, les évaluateurs n'ont donc pas de raisons stratégiques de modifier les notes (du moins si l'on suppose qu'ils ont confiance dans la confidentialité du système et dans sa finalité exclusivement développementale). Que constate-t-on alors ? La qualité du feed-back est meilleure, mais le niveau du bruit systémique reste élevé. Il représente toujours une plus grande part de la variance que la performance de la personne évaluée. En d'autres termes, même quand les évaluations sont faites avec des intentions purement développementales[9], et donc *a priori* sincères, il y reste beaucoup de bruit.

## *Un problème ancien, mais toujours pas résolu*

Si les systèmes d'évaluation de la performance sont si déficients, il semble logique que les organisations s'efforcent de les améliorer. Et en effet, depuis des dizaines d'années, d'innombrables réformes ont été entreprises dans ce domaine. Elles ont souvent employé certaines des stratégies de réduction du bruit que nous avons examinées. Mais il reste, à notre avis, beaucoup à faire.

La quasi-totalité des organisations utilisent d'une manière ou d'une autre ce que nous avons appelé *l'agrégation* de jugements multiples. En particulier, les systèmes d'évaluation à 360 degrés, qui agrègent les jugements de plusieurs évaluateurs, sont progressivement devenus la norme dans les grandes entreprises depuis les années 1990. (La revue *Human Resources Management* a publié un numéro spécial sur le feed-back à 360 degrés en 1993.)

Faire la moyenne des évaluations réalisées par plusieurs évaluateurs devrait bien sûr contribuer à réduire le bruit systémique, mais il est intéressant de noter que ce n'est pas pour résoudre ce problème-là que les systèmes de feed-back à 360 degrés ont été conçus. Leur principal objet est de mesurer d'autres aspects de la performance que ceux que voit habituellement une supérieure hiérarchique. Quand on demande à vos pairs et à vos subordonnés, et pas seulement à votre supérieure, de contribuer à l'évaluation de votre performance, la nature même de ce qui est évalué change. Et, en théorie, ce changement est nécessaire : pour exceller dans les organisations d'aujourd'hui, il ne suffit plus de satisfaire les attentes de son chef. Ce n'est pas par hasard que le développement du feed-back à 360 degrés a coïncidé avec la généralisation des organisations flexibles et des équipes projets.

Le feed-back à 360 degrés est un outil valide, au sens où – d'après certaines études – il est associé à des mesures de performance objectivement mesurables[10]. Malheureusement, son utilisation crée elle-même des difficultés. Comme rien n'est plus facile que d'allonger la liste des questions posées dans un questionnaire informatisé, et comme les entreprises ne cessent de multiplier les objectifs et les contraintes associés à chaque description de poste, beaucoup de questionnaires de feed-back sont d'une complexité qui frise l'absurde. Il est courant que leur longueur les rende impraticables[11]. Un exemple parmi d'autres : ce questionnaire qui compte quarante-six critères de notation, sur onze dimensions du poste, pour chaque évaluateur et chaque personne évaluée… Seul un évaluateur surhumain pourrait mobiliser le souvenir de faits exacts et pertinents sur autant d'aspects du travail de toutes les personnes qu'il évalue ! De telles usines à gaz ne sont pas seulement inutiles, mais aussi pernicieuses, car l'effet de halo fait que les différentes dimensions d'un poste, en principe séparées, ne seront pas traitées séparément. Une évaluation très positive ou

très négative sur une des premières questions posées aura tendance à tirer les réponses aux questions suivantes dans la même direction.

Le développement du feed-back à 360 degrés soulève un problème plus important encore : l'accroissement exponentiel du temps consacré à fournir du feed-back. Il n'est pas rare pour des cadres intermédiaires de devoir remplir des dizaines de questionnaires sur leurs collègues à tous les niveaux de l'organisation, voire sur leurs homologues dans d'autres entreprises, quand celles-ci demandent à leurs clients, fournisseurs ou partenaires de contribuer à l'évaluation de leur personnel. Aussi bien intentionnées que soient ces initiatives, elles produisent une explosion du nombre de questionnaires auxquels doivent répondre des évaluateurs déjà très occupés. Cela ne peut pas être bon pour la qualité des réponses. Les systèmes de feed-back à 360 degrés pourraient donc être un exemple de stratégie de réduction du bruit qui crée plus de coûts qu'elle ne génère de bénéfices – un problème sur lequel nous reviendrons dans la partie VI.

Enfin, les systèmes à 360 degrés n'échappent pas à un problème quasi universel dans les mesures de performance : l'inflation des notes. On a ainsi pu observer, dans une grande entreprise industrielle dont la performance globale n'était pas particulièrement impressionnante, que 98 % des cadres avaient été jugés « répondre pleinement aux attentes[12] ». Quand tout le monde (ou presque) reçoit des évaluations identiques et apparemment déconnectées de la réalité, il est permis de se demander à quoi elles servent.

## *Éloge du jugement relatif*

En théorie, il existe une solution efficace au problème de l'inflation des notes : la standardisation. Une pratique répandue pour ce faire est le *forced ranking* (classement forcé)[13]. Dans un système de classement forcé, les évaluateurs ne peuvent pas distribuer sans compter les bonnes notes : ils

sont obligés de se plier à une distribution préétablie. Jack Welch, quand il était P-DG de General Electric, était sans doute le plus fervent défenseur du *forced ranking*, considéré comme un moyen de mettre fin à l'inflation des évaluations et d'assurer la « sincérité » des revues de performance. Suivant son exemple, un grand nombre d'entreprises l'ont adopté, mais la plupart l'ont l'abandonné par la suite, après avoir constaté des effets néfastes sur le moral des employés et la qualité du travail en équipe.

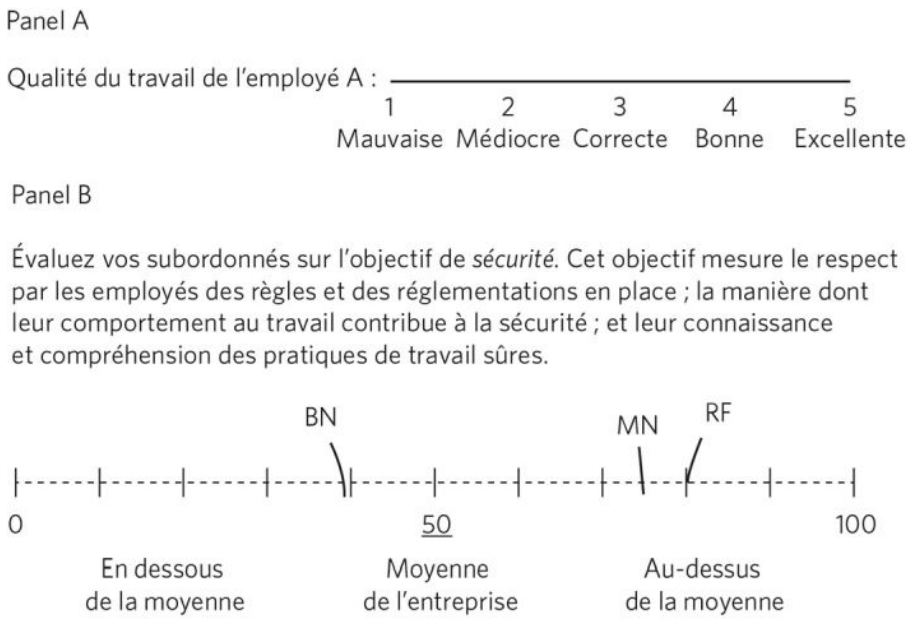



**Figure 17**[14]**.** Exemples d'échelles d'évaluation absolues et relatives.

Quels que soient par ailleurs leurs défauts, les classements, on l'a vu, sont moins sujets au bruit que les notations. L'exemple des dommages et intérêts punitifs nous a montré qu'il y a beaucoup moins de bruit dans les jugements relatifs que dans les jugements absolus, et cette observation s'applique aussi aux évaluations de performance[15].

Pour comprendre pourquoi, regardons la figure 17, qui donne deux exemples d'échelles d'évaluation de performance. Le panel A, dans lequel

un employé est évalué sur une échelle absolue, vous demande d'établir une équivalence, comme celles que nous avons décrites au chapitre 14 : il vous faut établir une équivalence entre l'impression que vous vous faites de la « qualité du travail » de l'employé et la note qui lui correspond le mieux. Dans le panel B, en revanche, chaque employé doit être comparé à un groupe d'autres employés sur un ensemble précis de comportements relatifs à la sécurité. L'évaluateur doit indiquer le classement (ou le positionnement relatif) de l'employé dans la population de référence, au moyen d'une échelle de percentiles. On voit sur le graphique que l'évaluateur a placé sur cette échelle trois employés, désignés par leurs initiales.

L'approche qu'illustre le panel B a deux avantages. Premièrement, évaluer les employés sur un seul critère à la fois (ici, la sécurité) est un bon exemple d'une stratégie de réduction du bruit dont nous parlerons davantage dans le prochain chapitre : la *structuration* d'un jugement complexe, décomposé en plusieurs critères. La structuration aide à limiter l'effet de halo, à condition que l'évaluation se fasse séparément sur chaque critère et que ces critères sont bien définis. Dans la figure 17, par exemple, on parle de « sécurité » et non d'un critère global et vague de « qualité du travail ».

Deuxièmement, comme nous l'avons vu au chapitre 15, un classement, c'est-à-dire un jugement relatif, réduit à la fois le bruit de pattern et le bruit de niveau. Vous aurez moins tendance à être incohérent (et donc à produire du bruit de pattern) si vous comparez la performance de deux membres de votre équipe que si vous donnez séparément une note à chacun d'eux. Mais surtout, les classements éliminent mécaniquement le bruit de niveau. Si Lynn est plus indulgente que Mary, quand Lynn et Mary évaluent le même groupe de vingt employés, leurs évaluations moyennes seront différentes. Mais leurs classements moyens seront identiques. Un évaluateur indulgent et un évaluateur intraitable donnent des notes différentes, mais utilisent les mêmes rangs.

Il n’est donc pas surprenant qu’on invoque souvent la réduction du bruit de niveau pour justifier le *forced ranking* : avec un classement forcé, les évaluations de tous les évaluateurs ont la même moyenne et la même distribution. Les classements sont « forcés » quand on stipule par avance la distribution des évaluations. La consigne peut être, par exemple, de ranger un maximum de 20 % des individus évalués dans la catégorie la plus élevée, et 15 % au moins dans la catégorie la plus basse. Comme cette consigne s’applique uniformément à tous les évaluateurs, le bruit de niveau est quasi éliminé.

## *Classez, mais sans forcer*

En principe, donc, le classement forcé devrait contribuer à réduire le bruit dans les évaluations, et donc à résoudre les problèmes que nous avons énumérés. Mais il a souvent de sérieux effets indésirables. Nous n’avons pas pour ambition de les passer tous en revue ici, d’autant qu’ils sont souvent liés à une mise en œuvre déficiente du système d’évaluation, plutôt qu’à son principe. Mais deux des problèmes que posent les systèmes de *forced ranking* sont des problèmes de fond, et, à ce titre, sont riches d’enseignements pour la réduction du bruit dans d’autres domaines.

Le premier de ces problèmes est la confusion entre performance absolue et performance relative. Il est certainement impossible que 98 % des cadres d’une entreprise fassent partie des meilleurs 20, 50 ou 80 % de leur groupe de pairs. Mais il n’est pas impossible que tous « répondent aux attentes » si ces attentes ont été définies *ex ante* et *en termes absolus*.

Un grand nombre de dirigeants d’entreprise contestent l’idée que la quasi-totalité des employés puissent « répondre aux attentes ». Si c’est le cas, disent-ils, c’est que les attentes en question ne sont pas assez exigeantes : des évaluations laxistes sont le reflet d’une culture d’entreprise

complaisante. Cette interprétation est parfois juste, bien sûr. Mais on ne peut pas exclure que la plupart des employés s'élèvent réellement à la hauteur d'attentes légitimement *élevées*. C'est d'ailleurs précisément ce qu'on s'attend à voir dans une organisation à très haute performance. On n'accuserait pas la NASA d'être laxiste dans ses procédures d'évaluation si elle jugeait, au retour d'une mission spatiale réussie, que tous ses astronautes ont pleinement répondu aux attentes.

La conclusion de tout cela est qu'un système de classement forcé n'a de sens que pour une organisation qui se soucie essentiellement de *performance relative*. C'est le cas en particulier quand, indépendamment de la performance absolue des individus, seul un pourcentage préétabli d'entre eux peut être distingué – comme lorsqu'on évalue des colonels dont certains seulement pourront être promus généraux. Il est parfaitement illogique, en revanche, de forcer une distribution prédéfinie des évaluations tout en prétendant mesurer un niveau *absolu* de performance – ce que font pourtant beaucoup d'entreprises. Et quand ces distributions prédéfinies imposent qu'un certain pourcentage d'employés soient notés comme inférieurs à des attentes prétendument absolues, elles ne sont pas seulement cruelles : elles sont absurdes. Il serait stupide de dire que 10 % des soldats d'une unité d'élite de l'armée doivent nécessairement être notés « insatisfaisants ».

Le second problème de principe que pose le classement forcé est que la distribution forcée qu'on impose aux évaluations est censée refléter la distribution des performances sous-jacentes réelles. On la suppose, en général, proche d'une distribution normale. Mais, même si l'on admet que la distribution des performances dans la population évaluée est conforme à ce postulat (ce qui est loin d'être certain), cette distribution risque fort de ne pas se retrouver dans une sous-population, comme le groupe des personnes notées par le même évaluateur. Si vous choisissez au hasard dix individus dans une population de plusieurs milliers, il n'est pas certain que deux d'entre eux exactement appartiendront aux meilleurs 20 % de la population

générale. (« Pas certain » est une litote : il n'y a que 30 % de chances que cela soit le cas.) En pratique, le problème est plus sérieux encore, car la composition des équipes n'est jamais aléatoire. Pour des raisons aisément imaginables, certaines équipes peuvent être composées presque entièrement d'employés très performants, tandis que dans d'autres, les employés en dessous de la moyenne sont surreprésentés.

Inévitablement, dans un tel contexte, un classement forcé sera source d'erreur et d'injustice. Supposons qu'un évaluateur doive noter cinq personnes dont les performances sont toutes très satisfaisantes et ne peuvent pas être distinguées les unes des autres. Forcer l'évaluateur à produire une distribution de notes différenciée là où la réalité est indifférenciée ne réduit pas l'erreur, mais, au contraire, l'accroît.

Les critiques du classement forcé ont souvent concentré leurs attaques sur le principe du classement lui-même, jugé brutal, inhumain et finalement contre-productif. Que vous acceptiez ou non ces arguments, le défaut fatal du classement forcé n'est pas d'être un « classement » : c'est d'être « forcé ». Chaque fois qu'on force un jugement à utiliser une échelle inadaptée, soit parce qu'on utilise une échelle relative pour mesurer une performance absolue, soit parce que les juges sont contraints à distinguer l'indistinguable, le choix de l'échelle ajoute automatiquement du bruit.

## *Et après ?*

Au regard des efforts faits par les organisations pour améliorer la mesure de la performance, c'est peu dire que les résultats ont été décevants. Car tous ces efforts ont aussi fait exploser le coût des processus d'évaluation. En 2015, par exemple, le cabinet d'audit et de conseil Deloitte estimait investir chaque année deux millions d'heures pour réaliser l'évaluation de ses soixante-cinq mille salariés[16]. Les évaluations de performance sont toujours

l’un des rituels organisationnels les plus redoutés, tant par ceux qui doivent les réaliser que par ceux qui s’y soumettent. Une enquête faite auprès de cadres, d’employés et de responsables des ressources humaines a montré que 90 % d’entre eux estiment que leurs processus de gestion de la performance ne produisent pas les résultats attendus[17].

La recherche confirme ce que la plupart des cadres vivent. Bien que le feed-back sur la performance, quand il est associé à un plan de développement pour l’employé, puisse apporter des améliorations, les évaluations, telles qu’elles sont le plus souvent pratiquées, démotivent autant qu’elles motivent. Comme le résume une étude, « en dépit de tout ce qui a été tenté pendant des dizaines d’années pour améliorer les processus [de gestion de la performance], ceux-ci continuent à générer des informations inexactes et ne sont pratiquement d’aucune utilité pour améliorer la performance[18] ».

En désespoir de cause, un nombre réduit mais croissant d’entreprises envisagent désormais une solution radicale : éliminer totalement les systèmes d’évaluation de la performance. Les partisans de cette « révolution dans la gestion de la performance[19] » sont essentiellement des entreprises des secteurs de la technologie et des services aux organisations, ainsi qu’une poignée d’acteurs issus de secteurs plus traditionnels. Leur objectif est de concentrer les systèmes de pilotage de la performance sur le feed-back développemental, tourné vers l’avenir, plutôt que sur l’évaluation des performances passées. Quelques-unes ont même adopté le principe des évaluations « sans chiffres », ce qui signifie qu’elles ont abandonné les traditionnelles notations.

Pour les entreprises qui n’y ont pas renoncé (c’est l’immense majorité), que peut-on faire pour les améliorer ? Une des principales stratégies de réduction du bruit repose sur le choix de la bonne échelle. L’objectif d’une échelle bien calibrée s’inspire du principe que nous avons appelé plus haut la « vision externe » : il est d’établir un *cadre de référence commun*. La

recherche montre que le choix de meilleurs formats d'évaluation et la formation des évaluateurs permettent d'obtenir plus de cohérence dans l'utilisation de l'échelle.

Pour cela, les échelles d'évaluation doivent tout d'abord être « ancrées » sur des descripteurs suffisamment spécifiques pour être interprétés de façon cohérente. Beaucoup d'organisations utilisent des *échelles comportementales ancrées* : chaque degré de ces échelles correspond à la description d'un comportement précis. Le panel de gauche de la figure 18 en donne un exemple.

La recherche montre, cependant, que les échelles comportementales ancrées ne suffisent pas à éliminer le bruit[20]. L'une des techniques qui permettent d'assurer une meilleure cohérence entre évaluateurs est de leur dispenser une formation spécifique, dite *formation au cadre de référence*. Dans une formation au cadre de référence, les évaluateurs apprennent à reconnaître les différentes dimensions de la performance et à en calibrer l'évaluation. Pour cela, ils visionnent des vidéos de cas pratiques, s'exercent à évaluer la performance qu'ils ont observée, puis comparent leur notation avec les évaluations « justes » fournies par des experts[21]. Le grand avantage de ces cas pratiques est qu'ils servent de référence : chaque cas définit un degré sur l'échelle de la performance. L'échelle d'évaluation devient une *échelle de cas*, comme dans le panel de droite de la figure 18.

Quand on utilise une échelle de cas, chaque évaluation d'un nouvel individu devient une comparaison : l'individu à évaluer est comparé avec ces « cas-ancres », que tout le monde connaît, et qui servent de référence partagée. L'évaluation est transformée en jugement relatif, en comparaison du cas présent avec les cas-ancres. Dans la figure 18, par exemple, Bob et Linda sont évalués par comparaison avec les cas-ancres de Jenny P. et Dan F. Comme il y a moins de bruit dans les jugements comparatifs que dans les évaluations absolues, les échelles de cas sont plus fiables que les échelles

graduées par des chiffres, par des adjectifs, ou même par des descriptions de comportements.

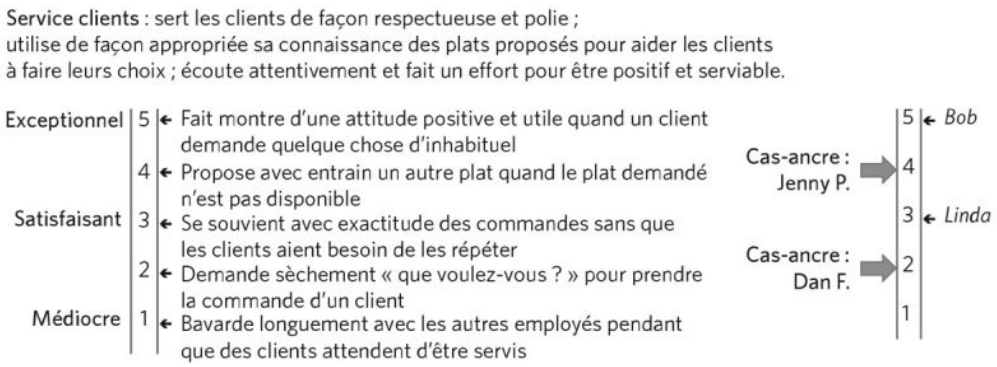



**Figure 18**[22]**.** Exemple d'une échelle comportementale ancrée (à gauche) et d'une échelle de cas (à droite).

Le principe de la formation au cadre de référence est connu depuis des décennies, et il est démontré que ces formations permettent de réaliser des évaluations plus exactes et moins sujettes au bruit. Elles sont pourtant loin de s'être généralisées. Il n'est pas difficile de deviner pourquoi. La formation au cadre de référence, la définition d'échelles de cas et les autres outils du même type sont complexes et prennent du temps. Pour être utiles, il faut qu'ils soient spécifiques, donc généralement conçus sur mesure pour une entreprise, voire pour une fonction ou un service. Les échelles et les cas-ancres doivent aussi être remis à jour fréquemment pour s'adapter à l'évolution des emplois et des attentes. En somme, ces outils exigent de l'entreprise qu'elle investisse encore davantage dans des systèmes de gestion de la performance qui lui coûtent déjà cher. La mode actuelle va dans le sens opposé. C'est un exemple de situation où le coût de la réduction du bruit est loin d'être nul : nous y reviendrons dans la partie VI.

En outre, toute organisation qui réduit le bruit attribuable aux évaluateurs limite aussi leur marge de manœuvre, et donc leur capacité

d’influencer les évaluations dans le sens de leurs propres objectifs. Leur demander de suivre une formation supplémentaire et de consacrer davantage d’efforts au processus d’évaluation, tout en renonçant à une partie de leur pouvoir d’influer sur les décisions, ne peut que susciter de fortes résistances. Il est d’ailleurs révélateur que la plupart des études sur la formation des évaluateurs au cadre de référence n’aient pas été réalisées sur des cadres d’entreprise, mais sur des étudiants[23].

---

Le sujet de l’évaluation de la performance soulève un grand nombre de questions, tant pratiques que philosophiques. Certains se demandent, par exemple, dans quelle mesure l’idée de performance individuelle a encore un sens dans les organisations d’aujourd’hui, où les résultats dépendent largement du travail d’équipe. Même si l’on convient que la performance individuelle a du sens, on peut s’interroger sur la distribution des niveaux de performance au sein d’une organisation : est-ce une distribution normale, par exemple, ou existe-t-il des « superstars[24] » qui contribuent de manière disproportionnée ? Enfin, si l’on cherche à faire en sorte que les gens donnent le meilleur d’eux-mêmes, il n’est pas déraisonnable de se demander si manier la carotte et le bâton sur la base de notes de performance individuelle est bien la meilleure méthode (ou même une méthode efficace).

Si vous avez pour tâche de concevoir ou de réviser un système de gestion de la performance, vous devrez traiter ces questions, et bien d’autres. Notre ambition ici n’est pas d’y répondre. Notre propos, plus modestement, est de suggérer une hypothèse : si vous faites des évaluations de performance, il y a de bonnes chances qu’elles se composent pour une très large part de bruit systémique, et qu’elles soient, pour cette raison, fondamentalement inutiles et même contre-productives.

Ce bruit dans les évaluations ne peut pas être éliminé par de simples solutions technologiques. Pour s’y attaquer, il vous faudra réfléchir

clairement aux jugements que les évaluateurs sont censés produire. Vous vous rendrez très probablement compte que ces jugements peuvent être améliorés en clarifiant l'échelle d'évaluation et en apprenant aux évaluateurs à s'en servir de façon cohérente. Cette stratégie de réduction du bruit – la définition d'une meilleure échelle – est applicable dans de nombreux autres domaines.

### À propos de la définition de l'échelle

« Nous consacrons beaucoup de temps à nos évaluations de performance, et pourtant, ce que nous mesurons, c'est un quart de performance, trois quarts de bruit systémique. »

« Nous avons essayé le feed-back à 360 degrés et le classement forcé pour résoudre ce problème, mais nous avons peut-être aggravé les choses. »

« S'il y a autant de bruit de niveau, c'est parce que les différents évaluateurs ont des idées complètement différentes sur ce que signifient les mots "bien" ou "très bien". Ils ne se mettront d'accord que si nous leur donnons des cas concrets pouvant servir d'ancres sur l'échelle d'évaluation. »

CHAPITRE 24

# Recrutement : la structuration des jugements complexes

Si vous avez déjà postulé à un emploi, les mots *entretien d'embauche* évoquent sans doute un souvenir vivace, et probablement désagréable. L'entretien d'embauche, durant lequel un candidat rencontre un futur supérieur hiérarchique ou un professionnel des ressources humaines, est un rite de passage obligé pour entrer dans la plupart des organisations.

En général, l'entretien suit une routine bien huilée. Après les politesses d'usage, le recruteur demande à la candidate de raconter son expérience professionnelle et d'en développer certains aspects. Les questions portent sur ses réussites et sur les difficultés qu'elle a rencontrées, sur sa motivation pour le poste, ou sur ses suggestions d'amélioration pour l'entreprise. On demandera souvent à la candidate de décrire sa personnalité et d'expliquer pourquoi elle est particulièrement adaptée au poste ou à la culture maison. Parfois, on parlera de ses hobbies ou de ses centres d'intérêt extraprofessionnels. Vers la fin de l'entretien, la candidate sera souvent invitée à poser à son tour quelques questions, qui seront dûment évaluées pour leur pertinence et leur perspicacité.

Si vous êtes vous-même dans la position du recruteur, vos méthodes se rapprochent probablement de ce rituel. Comme l'écrivait un psychologue

du travail et des organisations, « il est rare, voire inconcevable, que quelqu'un soit recruté sans passer par un entretien sous une forme ou sous une autre[1] ». Et la quasi-totalité des professionnels qui prennent des décisions de recrutement à l'issue de ces entretiens se reposent à un degré ou à un autre sur les jugements intuitifs qu'ils y ont formés[2].

L'universalité de l'entretien d'embauche reflète une croyance profondément enracinée : quand il s'agit de choisir les personnes avec lesquelles nous allons travailler, rien ne vaut le jugement. Or, cet exercice de jugement a une particularité qui nous intéresse beaucoup ici : du fait de son universalité et de son importance, les psychologues du travail et des organisations l'ont étudié de manière approfondie. Dès 1917, le premier numéro du *Journal of Applied Psychology* parlait de la sélection du personnel comme du « problème suprême […] car les capacités humaines sont, en définitive, les principales ressources de la nation[3] ». Un siècle plus tard, nous en savons beaucoup plus sur l'efficacité des différentes techniques de recrutement (dont l'entretien traditionnel). Aucun travail de jugement complexe n'a fait l'objet d'autant d'études de terrain. Cela en fait un sujet exemplaire, dont les enseignements peuvent être extrapolés à beaucoup de jugements où l'on doit choisir, dans un contexte d'incertitude, entre plusieurs options.

## *Les pièges de l'entretien d'embauche*

Si vous ne connaissez pas les travaux de recherche sur l'entretien d'embauche, ce qui suit va sans doute vous surprendre. En substance, si votre objectif est de déterminer si des candidats vont réussir ou échouer dans un emploi donné, l'entretien classique (dit « non structuré », pour le distinguer de l'entretien structuré, sur lequel nous reviendrons) n'est pas

très informatif. Pour le dire crûment, il est même souvent complètement inutile.

Cette conclusion provient d'innombrables études sur la qualité du jugement de recrutement. Leur méthode habituelle consiste à estimer la corrélation entre deux données : d'une part l'évaluation qu'effectue un évaluateur après un entretien, d'autre part la réussite du candidat une fois embauché. Si la corrélation entre évaluation en entretien et réussite dans le poste est élevée, alors l'entretien est un bon prédicteur de la performance au travail. Inversement, si la corrélation est faible, c'est que l'entretien ne permet pas de prédire la performance. Il en va de même pour toute autre technique de sélection, pour laquelle on calcule de la même manière la corrélation avec la réussite.

Une précision est ici nécessaire. Comme le chapitre précédent l'a souligné, la définition de la réussite n'est pas un problème simple. Dans les études dont nous parlons ici, la performance au travail est généralement quantifiée par les évaluations d'un supérieur hiérarchique, ou parfois en prenant en compte la durée pendant laquelle la personne recrutée reste en poste. La validité de ces mesures est bien sûr, on l'a vu, loin d'être parfaite. Toutefois, s'agissant d'évaluer la qualité des jugements d'une organisation qui sélectionne des employés pour leur aptitude à tenir un poste, il semble raisonnable d'utiliser les jugements produits par la même organisation lorsqu'elle évalue leur performance dans ce poste. Toute analyse de la qualité des décisions de recrutement doit nécessairement faire cette hypothèse.

Or que concluent ces analyses ? Au chapitre 11, nous avons mentionné une étude faisant état d'une corrélation de 0,28 entre les notes d'entretiens et les évaluations de performance. D'autres études estiment des corrélations allant de 0,20 à 0,33[4]. Comme nous l'avons vu, ce niveau de corrélation est très satisfaisant pour des chercheurs en sciences sociales, mais pas très utile pour prendre des décisions. En utilisant le pourcentage de concordance

(PC), on peut calculer une probabilité : au vu des niveaux précédents de corrélation, si tout ce que vous savez de deux candidates est qu'au cours de l'entretien l'une des deux vous a parue meilleure que l'autre, les chances que cette candidate soit effectivement la plus performante une fois embauchée vont de 56 à 61 %. C'est un peu mieux que de choisir la candidate en jouant à pile ou face, mais c'est loin d'être un moyen très fiable de prendre des décisions importantes.

Il est vrai que l'entretien d'embauche ne sert pas seulement à produire un jugement sur un candidat. Il est aussi l'occasion de faire la promotion de l'entreprise auprès des candidats prometteurs et de commencer à bâtir une relation de travail avec d'éventuels futurs collègues. Mais du point de vue d'une organisation qui consacre du temps et des efforts à choisir les meilleurs talents, le principal objet de l'entretien est clairement la sélection. Et à cet égard, ce n'est pas une grande réussite.

## *Du bruit dans l'entretien*

On comprend aisément pourquoi l'entretien traditionnel produit de l'erreur dans la prévision d'une performance future. Tout d'abord, une part importante de cette erreur est liée à ce que nous appelons l'ignorance objective[5] (voir chapitre 11). La performance au travail dépend de nombreux facteurs : la rapidité avec laquelle la nouvelle recrue s'adapte à son poste, la manière dont les différents événements de la vie affectent son travail… Tous ces facteurs sont largement imprévisibles au moment de l'embauche. Cette incertitude limite non seulement la validité prédictive de l'entretien, mais aussi celle de toutes les autres techniques de sélection.

L'entretien d'embauche est aussi une mine de biais psychologiques. Depuis quelques années, presque tous les dirigeants ont entendu parler des études qui font état de biais dans le recrutement : lorsqu'ils font passer des

entretiens d'embauche, les recruteurs ont tendance, souvent sans en être conscients, à favoriser les candidats qui leur ressemblent culturellement, ou ceux avec lesquels ils ont quelque chose en commun, notamment le genre, l'appartenance ethnique ou le cursus éducatif[6]. Un grand nombre d'entreprises reconnaissent désormais les risques engendrés par ces biais et tentent, avec plus ou moins de succès, de les maîtriser en dispensant des formations spécifiques aux professionnels du recrutement et à leurs autres responsables. On sait par exemple que l'apparence physique joue un grand rôle dans l'évaluation des candidats, même pour des postes où elle ne devrait peser que très peu ou pas du tout. Ce type de biais, parce qu'il est commun à la plupart des recruteurs, tend à produire une erreur partagée – un biais négatif ou positif dans l'évaluation d'un candidat donné.

Vous ne serez pas étonné d'apprendre qu'il y a aussi du bruit dans les entretiens. Différents recruteurs réagissent différemment au même candidat et arrivent à des conclusions différentes. Les mesures de la corrélation entre les évaluations produites par deux recruteurs après un entretien avec le même candidat se situent entre 0,37 et 0,44 (PC = 62 à 65 %)[7]. Cela peut s'expliquer par le fait que, quand un candidat passe deux entretiens successifs, il ne se comporte pas exactement de la même façon avec les deux recruteurs. Cependant, même quand l'entretien est réalisé avec un jury, ce qui garantit que les membres du jury observent exactement le même comportement, la corrélation entre leurs évaluations est encore loin d'être parfaite. Une méta-analyse l'estime à 0,74 (PC = 76 %). En d'autres termes, lorsque vous et un collègue ferez partie du même jury et interviewerez ensemble les deux mêmes candidats, vous serez encore en désaccord une fois sur quatre pour savoir lequel est le meilleur des deux.

Ces différences reflètent, pour une bonne part, du bruit de pattern : chaque recruteur a ses propres préférences et a donc une réaction différente aux mêmes candidats. La plupart des grandes organisations sont conscientes de ce problème et ont pour règle que le même candidat soit reçu par

plusieurs recruteurs, afin d'agréger ensuite les résultats. Malheureusement, l'opinion « agrégée » se forme en général à l'issue d'une discussion qui vise à dégager un consensus, une procédure qui crée ses propres problèmes, comme nous l'avons déjà noté.

Résultat plus inattendu, les entretiens d'embauche sont largement sujets au bruit occasionnel. Nombre d'études montrent ainsi, par exemple, que les décisions de recrutement sont très liées aux premières impressions qui se forment au cours de la phase informelle de prise de contact, ces deux ou trois premières minutes de conversation informelle pendant lesquelles le recruteur bavarde aimablement avec le candidat pour le mettre à l'aise avant d'entrer dans le vif du sujet. Ces premières impressions s'avèrent avoir beaucoup d'importance[8].

Peut-être pensez-vous que juger quelqu'un sur votre première impression n'est pas un problème. De fait, nos premières impressions sont, au moins en partie, significatives. Nous savons tous que les premières secondes d'une interaction avec un inconnu nous apprennent quelque chose sur lui. On pourrait penser que ce qui est vrai pour n'importe lequel d'entre nous l'est plus encore pour des recruteurs chevronnés, et que ceux-ci devraient donc être capables de former des jugements instinctifs, rapides, et justes. Hélas, il apparaît que ce qu'on évalue dans les premières secondes d'un entretien, ce sont précisément les qualités superficielles associées aux premières impressions : principalement l'extraversion et les compétences verbales d'un candidat. Même la qualité d'une poignée de main est un prédicteur significatif des recommandations de recrutement[9] ! Une solide poignée de main est sans doute un atout appréciable, mais combien de recruteurs choisiraient-ils consciemment d'en faire un critère de recrutement ?

## *La psychologie du recruteur*

Pourquoi les premières impressions finissent-elles par influencer le résultat d'un entretien qui dure bien plus longtemps ? D'abord parce que la recruteuse, dans l'entretien traditionnel, est libre d'orienter la discussion dans le sens qu'elle juge bon. Elle va donc souvent poser des questions qui tendent à confirmer son impression première. Si un candidat lui paraît timide et réservé, par exemple, la recruteuse ne manquera pas de le tester en lui posant des questions directes (et difficiles) sur son expérience du travail en équipe ou de la prise de parole en public. Une candidate qui lui a semblé sociable et enjouée ne sera pas soumise au même interrogatoire. Cela peut certes sembler logique, mais le résultat est que les éléments recueillis sur ces deux candidats ne seront pas comparables. Une étude a ainsi montré que la première impression des recruteurs, fondée sur les CV des candidats et leurs résultats à des tests écrits, avait un effet considérable sur le déroulement de l'entretien (et ce, que la première impression soit positive ou négative). Les recruteurs dont la première impression est positive, par exemple, posent moins de questions et ont tendance à utiliser une partie de l'entretien pour commencer à « vendre » l'entreprise au candidat[10].

La force des premières impressions n'est pas le seul écueil à éviter. Un autre problème, dans les entretiens, est que les recruteurs veulent que le candidat assis devant eux *fasse sens* : c'est une manifestation de ce que nous avons décrit au chapitre 13 comme notre tendance à l'excès de cohérence. Une étude l'a démontré d'une manière assez stupéfiante. Les chercheurs ont demandé à des étudiants de jouer soit le rôle du recruteur, soit celui du candidat, en leur imposant simplement que l'entretien n'utilise que des questions fermées, c'est-à-dire auxquelles on puisse répondre par oui ou par non[11]. Certains des étudiants qui jouaient le rôle de candidats ont ensuite reçu l'instruction de répondre aux questions *de façon aléatoire*. (Leur réponse, oui ou non, était déterminée par la première lettre de la question posée.) Comme l'écrivent ironiquement les chercheurs : « Certains étudiants-candidats craignaient de prime abord que l'absurdité de leurs

réponses aléatoires ne saute aux yeux et que l'exercice ne doive être interrompu. Mais il n'en a rien été, et les entretiens se sont tous déroulés normalement. » Vous avez bien lu : *pas un seul* des étudiants-recruteurs ne s'est rendu compte qu'on lui répondait au hasard. Pire encore : quand on a demandé à ces recruteurs d'estimer s'ils étaient « capables de tirer beaucoup de conclusions sur cette personne eu égard au peu de temps passé ensemble », ceux qui avaient été confrontés à des réponses aléatoires étaient aussi nombreux à répondre positivement que ceux auxquels les candidats avaient répondu sincèrement. Ainsi fonctionne notre étonnante aptitude à trouver de la cohérence là où il n'y en a pas. De même que nous trouvons souvent un pattern imaginaire dans des données aléatoires ou une forme dans les contours d'un nuage, nous sommes capables de trouver une logique à un ensemble de réponses parfaitement absurdes.

Pour prendre un exemple moins extrême et plus réaliste, considérons le cas suivant. Un jour, un des auteurs de ce livre faisait passer un entretien à un candidat qui, dans son précédent emploi, était directeur financier d'une entreprise moyenne. Observant que le candidat avait quitté ce poste au bout de quelques mois, il lui demanda pourquoi. Le candidat répondit qu'il avait démissionné en raison d'un « désaccord stratégique » avec le P-DG. Une heure plus tard, un collègue qui recevait à son tour le candidat en entretien lui posa la même question et obtint la même réponse. Mais dans la discussion de débriefing qui suivit, les deux interviewers furent d'opinions totalement différentes : l'un, qui avait eu une première impression positive du candidat, considérait sa décision de quitter l'entreprise comme un signe de courage et d'intégrité ; l'autre, dont la première impression avait été négative, y voyait un signe de rigidité, voire d'immaturité. L'anecdote est révélatrice : même si nous aimons croire que notre jugement sur un candidat repose sur des faits, le sens que nous donnons à ces faits est largement influencé par nos attitudes antérieures.

Les limites de l’entretien traditionnel jettent un doute sérieux sur la qualité des conclusions que nous pouvons en tirer. Pourtant, les impressions qui se forment au cours d’un entretien sont vivaces, et l’interviewer leur accorde généralement un grand crédit. Lorsque nous disposons, pour prendre une décision sur un candidat, à la fois d’un entretien et d’un ensemble d’autres informations, nous avons tendance à donner trop d’importance à l’entretien et trop peu à d’autres éléments également prédictifs, mais qui sont plus abstraits, comme les résultats d’un test.

Une autre anecdote, vécue par un autre des auteurs de ce livre, illustre cette observation. On demande souvent aux professeurs qui postulent pour un poste à l’université de simuler un cours devant un auditoire composé d’autres professeurs, afin de s’assurer que leurs compétences pédagogiques soient à la hauteur des exigences de l’institution. Les enjeux de cet exercice pour le postulant sont bien sûr beaucoup plus élevés que dans un cours ordinaire. Un jour où l’un d’entre nous assistait à une leçon de ce type, il nota la prestation médiocre du candidat, mais remarqua aussi que celle-ci ne pouvait s’expliquer que par la situation de stress où se trouvait le professeur : celui-ci, qui enseignait dans une université de premier plan, faisait état dans son CV de scores exceptionnels dans ses évaluations par les étudiants, ainsi que de plusieurs prix récompensant son excellence pédagogique. Pourtant, la vive impression négative produite par son échec dans une situation artificielle devait peser davantage dans la décision finale que des données objectivement plus fiables, mais abstraites, sur ses qualités pédagogiques.

Cet exemple illustre aussi une difficulté plus générale : quand l’entretien n’est pas la seule source d’information sur des candidats, parce qu’on dispose par exemple de tests, de références, ou d’autres éléments encore, il faut combiner ces données pour produire un jugement d’ensemble. Vous reconnaissez sans doute la question qui se pose alors : pour réaliser cette combinaison, faut-il utiliser un jugement (et réaliser ce

que nous avons appelé une prédiction clinique) ou une formule (et donc une prévision mécanique) ? Comme nous l'avons vu au chapitre 9, l'approche mécanique est supérieure en général, et le cas de la prédiction de la performance au travail ne fait pas exception. Malheureusement, les études montrent que l'immense majorité des professionnels des ressources humaines privilégient l'approche clinique[12]. Cette pratique ajoute une source de bruit supplémentaire à un processus qui, on l'a vu, n'en manque pas.

## *L'effet de la structuration sur la qualité de la sélection*

Si l'entretien traditionnel et le jugement clinique ont une validité prédictive limitée pour prendre des décisions d'embauche, comment peut-on en prendre de meilleures ? Fort heureusement, la recherche suggère aussi des réponses à cette question, et certaines entreprises n'y sont pas restées sourdes.

Parmi celles-ci, on trouve notamment Google, comme le raconte Laszlo Bock, qui en dirigea les ressources humaines, dans son livre *Work Rules !*. Son constat de départ : l'entreprise, en dépit de sa volonté affichée d'attirer des talents exceptionnels, et malgré les ressources considérables qu'elle consacrait au recrutement, était à la peine. Un audit de la validité prédictive des entretiens d'embauche conclut à « une absence totale de lien statistique […], des résultats complètement aléatoires[13] ». Les changements mis en place par Bock pour remédier à cette situation s'inspirent des travaux de recherche sur la sélection du personnel conduits depuis des décennies. Ils illustrent aussi plusieurs techniques d'hygiène de la décision.

Une de ces techniques devrait désormais vous être familière : l'agrégation. Son emploi dans ce contexte n'a rien d'étonnant. Presque toutes les entreprises agrègent les jugements de plusieurs recruteurs sur le même candidat. Google, où l'on ne fait pas les choses à moitié, faisait parfois passer jusqu'à vingt-cinq entretiens aux mêmes candidats ! Une des décisions de Bock fut de réduire ce nombre à quatre : les données lui montrèrent en effet que des entretiens supplémentaires n'ajoutaient presque rien à la validité prédictive des quatre premiers. Mais pour assurer ce niveau de validité, Bock mit en place une règle que toutes les entreprises n'appliquent pas : les recruteurs notent le candidat de façon séparée, *avant* de communiquer entre eux. Une fois encore, l'agrégation est une manière efficace de réduire le bruit, à condition toutefois que les jugements soient réellement indépendants.

Mais la principale inspiration de Google est une technique d'hygiène de la décision que nous n'avons pas encore décrite en détail : la *structuration des jugements complexes*. « Structurer » peut vouloir dire beaucoup de choses. Au sens où nous l'entendons ici, *structurer* un jugement complexe, c'est appliquer trois principes. D'abord, la *décomposition* d'un jugement en éléments séparés. Ensuite, l'*indépendance*, qui impose de gérer le processus de collecte d'informations pour s'assurer que les inputs soient indépendants les uns des autres. Enfin, le *jugement global différé*, qui retarde la discussion d'ensemble et le jugement final jusqu'à ce que tous ces inputs aient été recueillis.

Le premier principe, la *décomposition*, découpe le jugement en composantes, les *évaluations intermédiaires*. Le but est le même que lorsqu'une directive identifie les sous-jugements à réaliser : il s'agit de focaliser l'attention des juges sur les prédicteurs les plus importants. La décomposition sert de mode d'emploi, en indiquant les données qui sont nécessaires, et en écartant celles qui n'ont pas d'importance.

Dans l'exemple du recrutement chez Google, il y a quatre évaluations intermédiaires à faire sur chaque candidat : la capacité cognitive générale, le leadership, la compatibilité avec la culture de l'entreprise, et les compétences spécifiques au poste à pourvoir. (Certaines de ces évaluations sont à leur tour décomposées en évaluations intermédiaires plus fines.) Il est important, au passage, de noter ce qui n'apparaît *pas* sur cette liste : l'apparence physique séduisante d'un candidat, sa faconde, l'originalité de ses hobbies, ou tous les autres éléments positifs ou négatifs qu'un recruteur pourrait remarquer au cours d'un entretien non structuré, sont hors sujet.

On pourrait penser que définir ce type de structure avant de recruter relève du simple bon sens. Et de fait, quand on recrute un comptable débutant ou un assistant administratif, il existe des descriptions de poste standards qui détaillent les compétences nécessaires. Mais comme le savent tous les professionnels du recrutement, plus les postes s'élèvent en niveau de responsabilité ou sortent de l'ordinaire, et plus il devient difficile de spécifier précisément les compétences et qualités requises. C'est pourquoi cette étape indispensable est souvent occultée.

Un chasseur de têtes réputé souligne ainsi à quel point il est important de définir les compétences requises de façon suffisamment précise[14], ce qu'il appelle « investir dans la définition du problème » : au tout début du processus, *avant* de commencer à rencontrer des candidats, les décideurs doivent consacrer un temps suffisant à se mettre d'accord sur une description de poste claire, détaillée et sélective. Le risque à éviter est de produire, à force de consultations et de compromis, des descriptions de poste aussi vagues qu'interminables, qui listent tous les attributs du candidat idéal, mais qui n'en donnent ni hiérarchie ni pondération.

Le deuxième principe du jugement structuré, l'*indépendance*, veut que les informations qui contribuent à chaque évaluation intermédiaire soient recueillies de façon aussi séparée que possible. Car lister les éléments de la description de poste ne suffit pas à réduire le bruit : après tout, dans les

entretiens traditionnels, beaucoup de recruteurs savent aussi quelles sont les quatre ou cinq choses qu'ils cherchent à évaluer. Le problème est qu'ils n'évaluent pas ces éléments *séparément* au cours de l'entretien. Chaque évaluation influence les autres, ce qui y introduit beaucoup de bruit.

Pour remédier à ce problème, il faut des évaluations fondées sur des faits et indépendantes les unes des autres. La technique la plus importante pour s'en assurer (utilisée par beaucoup d'autres entreprises) est l'*entretien comportemental structuré*[15]. Dans ce type d'entretien, la tâche principale des recruteurs n'est pas de formuler une appréciation globale sur le candidat : elle est de recueillir des données sur chacun des critères d'évaluation, et de noter le candidat séparément sur chacun de ces critères. Pour ce faire, au cours d'un entretien structuré, les recruteurs posent au candidat des questions prédéfinies sur ses comportements passés dans diverses situations. Ils doivent aussi donner un score aux réponses du candidat, en se fondant sur une grille d'évaluation elle aussi prédéterminée. Pour chaque question, la grille indique des exemples de ce à quoi ressemblent des réponses moyennes, bonnes ou très bonnes. C'est donc un exemple d'échelle comportementale ancrée, comme celles que nous avons évoquées dans le chapitre précédent : la grille d'évaluation réduit le bruit dans le jugement de chacune des réponses à chacune des questions qui concourent à l'évaluation sur chacun des critères.

Si cette approche vous semble assez différente des conversations informelles qu'on peut avoir dans un entretien traditionnel, vous avez raison. En vérité, beaucoup d'entretiens structurés ressemblent davantage à un examen oral ou à un interrogatoire qu'à une rencontre professionnelle. C'est pourquoi ni les recruteurs, ni les candidats n'apprécient beaucoup les entretiens structurés (en tout cas, pas autant que les entretiens non structurés).

Pourtant, les données disponibles plaident sans ambiguïté en leur faveur. Les spécialistes du sujet ne sont pas toujours d'accord sur ce qui

définit un entretien structuré[16], mais quelle que soit la définition retenue, la littérature sur la sélection du personnel conclut de manière constante que l'entretien structuré prédit mieux la performance future que l'entretien traditionnel non structuré[17]. Les corrélations avec la performance future vont de 0,44 à 0,57. Traduits en pourcentage de concordance, ces chiffres suggèrent que vos chances de sélectionner le meilleur candidat avec un entretien structuré se situent entre 65 et 69 %, à comparer avec les 56 à 61 % que vous donnerait un entretien traditionnel.

En plus des entretiens structurés, on peut bien sûr recourir à d'autres techniques pour évaluer les candidats. Pour tester les compétences spécifiques à un poste, on peut ainsi utiliser des *tests sur échantillons de travail*[18], par exemple en demandant à un candidat programmeur d'écrire quelques lignes de code informatique. La recherche a montré que ce type de test fait partie des meilleurs prédicteurs de la performance au travail. (Ce n'est guère surprenant : pour savoir si quelqu'un peut s'acquitter d'une tâche, lui demander de l'effectuer et mesurer la qualité du résultat semble une méthode assez sensée.) Bock rapporte que Google utilise aussi des « références détournées » (*backdoor references*), fournies non pas par une personne désignée par le candidat, mais par des employés de Google dont il a croisé la route.

Le troisième principe du jugement structuré, le *jugement global différé*, se résume à une idée simple : ne pas bannir l'intuition, mais la garder pour la fin. Chez Google, comme dans beaucoup d'autres entreprises, la recommandation finale d'embauche est produite de façon collégiale par un comité de recrutement. Celui-ci analyse toutes les notes obtenues par les candidats sur chaque critère d'évaluation intermédiaire lors de chaque entretien, ainsi que toutes les autres informations disponibles. Sur cette base, le comité décide ou non de faire une proposition.

Cette décision d'embaucher n'est donc *pas* produite mécaniquement par l'agrégation des données collectées. Le paradoxe mérite d'être souligné :

Google, à l'évidence, valorise les données quantitatives ; et ses dirigeants connaissent les études démontrant la supériorité d'une combinaison mécanique de données sur une combinaison clinique. Pourtant, en dernière analyse, la décision d'embaucher y reste un jugement. Le comité de recrutement prend en considération tous les éléments recueillis, avant de se lancer dans une discussion sur la question : « Cette personne réussira-t-elle chez Google ? » La décision ne résulte pas d'un simple calcul.

Dans le prochain chapitre, nous expliquerons pourquoi nous pensons que cette approche de la décision finale est tout à fait raisonnable. Il faut observer cependant que si la décision d'embauche chez Google n'est pas mécanique, elle est tout de même ancrée sur la note moyenne donnée par les quatre recruteurs qui ont fait passer les entretiens. Elle est également éclairée par les autres éléments d'information qui sous-tendent ces notes. Autrement dit, Google permet au jugement et à l'intuition d'intervenir dans son processus de décision, mais seulement *après* que tous ces éléments ont été recueillis et analysés. De cette façon, la tendance de tout membre du comité de recrutement à se former une impression rapide et intuitive et à s'y tenir par la suite est quelque peu neutralisée.

Ces trois principes – la décomposition, l'évaluation indépendante et le jugement global différé – ne sont pas nécessairement un modèle pour toutes les organisations qui veulent améliorer leur processus de sélection. Mais ils sont cohérents avec les recommandations que les psychologues du travail et des organisations ont formulées au fil des années. Ces principes ressemblent d'ailleurs à la méthode de sélection que l'un d'entre nous (Kahneman) a mise en place dans l'armée israélienne dès 1956[19]. Cette méthode formalisait les critères d'évaluation (une liste de critères de personnalité et de compétence à évaluer). Elle exigeait des recruteurs qu'ils recueillent des données objectives sur chacun des critères, puis qu'ils les notent l'un après l'autre. Elle permettait, enfin, aux recruteurs de recourir à leur jugement

intuitif pour prendre leur décision finale – mais seulement après avoir déroulé les étapes de cette évaluation structurée.

---

Il existe une masse imposante de données scientifiques qui démontrent la supériorité des méthodes de jugement structuré (et notamment des entretiens structurés). Il existe également des conseils pratiques et aisément accessibles pour guider les cadres qui souhaitent adopter ces méthodes[20]. Il faut aussi souligner, comme l'ont observé à la fois les entreprises et les chercheurs qui ont mis en place ces méthodes, que le jugement structuré n'est pas plus coûteux, au contraire : peu de choses coûtent aussi cher que le temps consacré par des dirigeants à rencontrer des candidats.

La plupart des cadres dirigeants restent cependant convaincus de la valeur irremplaçable des méthodes informelles reposant sur les entretiens traditionnels. De façon remarquable, c'est aussi le cas de nombreux candidats à l'embauche, persuadés que seul l'entretien en face à face leur permettra de démontrer à un employeur leurs véritables qualités. Des chercheurs qui ont étudié ces croyances ont donné à leur article un sous-titre évocateur : « la persistance d'une illusion[21] ». S'il est une illusion persistante, en tout cas, c'est bien celle-ci : les recruteurs, comme les candidats, sous-estiment gravement le bruit qui pollue leurs jugements en matière d'embauche.

## À propos de la structuration des jugements

« Dans l'entretien informel traditionnel, nous avons souvent le sentiment intuitif et irrésistible de comprendre le candidat et de savoir immédiatement si la personne conviendra pour le poste. Nous devons apprendre à nous méfier de ce sentiment. »

« L'entretien traditionnel est dangereux non seulement en raison des biais qui l'affectent, mais aussi à cause du bruit auquel il est sujet. »

« Nous devons mettre de la structure dans nos entretiens et, plus largement, dans nos processus de sélection. Commençons par définir de façon plus claire et plus précise ce que nous recherchons chez les candidats. Assurons-nous qu'ils soient évalués de façon indépendante sur chacun des critères retenus. Enfin, différons le moment du jugement global. »

CHAPITRE 25

# Le protocole d’évaluations intermédiaires

Il y a quelque temps, deux d’entre nous (Kahneman et Sibony, avec notre ami Dan Lovallo) ont suggéré une méthode de prise de décision destinée aux organisations. Nous avons appelé cette méthode, dont la finalité principale est de réduire le bruit, le *protocole d’évaluations intermédiaires*[1]. Ce protocole intègre plusieurs des stratégies d’hygiène de la décision évoquées dans les chapitres précédents. Il est applicable à tout jugement qui consiste à examiner des plans ou options possibles, et qui nécessite de prendre en compte plusieurs dimensions d’évaluation. Il peut être utilisé et adapté par des organisations de toutes sortes – entreprises, universités ou agences gouvernementales.

Nous illustrerons ce protocole par un exemple schématique qui s’inspire de plusieurs cas réels, mettant en scène une entreprise fictive que nous appellerons Mapco. Mapco étudie la possibilité de faire une acquisition stratégique importante. Nous allons suivre pas à pas les étapes de sa prise de décision, et montrer de quelle manière ces étapes sont différentes de la démarche habituelle d’une entreprise dans cette situation. Comme vous le verrez, ces différences sont importantes, mais subtiles : peut-être même, pour un observateur inattentif, n’y aurait-il rien qui sorte de l’ordinaire.

## *Première réunion : se mettre d'accord sur l'approche*

L'idée d'acquérir Roadco, un concurrent important, était dans l'air depuis longtemps chez Mapco. La possibilité avait suffisamment mûri pour que les dirigeants de l'entreprise envisagent de soumettre cette proposition à leur conseil d'administration. Joan Morrison, la P-DG de Mapco, réunit donc le comité stratégique du conseil pour une discussion préliminaire sur cette possible acquisition. Mais elle orienta aussitôt la conversation sur la manière de s'assurer que les délibérations du conseil à ce sujet soient de bonne qualité. En particulier, dès le début de la réunion, elle surprit les membres du comité en faisant cette proposition :

« J'aimerais vous proposer une nouvelle procédure pour la réunion du conseil d'administration où nous prendrons une décision sur l'acquisition de Roadco. Il s'agit du protocole d'évaluations intermédiaires. Le nom est technocratique, mais l'idée est très simple. Elle s'inspire de la similitude entre l'évaluation d'une option stratégique et l'évaluation d'un candidat à l'embauche.

» Vous connaissez les études qui montrent que, dans le recrutement, les entretiens structurés produisent de meilleurs résultats que les entretiens non structurés, et, plus largement, que structurer une décision d'embauche permet de l'améliorer. Vous savez d'ailleurs que notre direction des ressources humaines a adopté ces principes pour le recrutement. Notre expérience confirme ce que la recherche a démontré : introduire de la structure nous conduit à des décisions plus exactes ; c'est une amélioration considérable par rapport aux entretiens non structurés que nous avions l'habitude de pratiquer.

» Pour moi, il y a une similitude frappante entre l'évaluation des candidats et celle des différentes options stratégiques qui s'offrent à nous. Pour le dire simplement : *les options sont comme des candidats*. Cette

similitude m’amène à penser que la méthode qui fonctionne pour évaluer des candidats fonctionnera aussi, en l’adaptant bien sûr, pour évaluer des options stratégiques. »

L’analogie proposée par Joan déconcerta d’abord les membres du comité. Certains répliquèrent en soulignant les différences entre recrutement et décision stratégique. Le processus de recrutement, expliquèrent-ils, est une machine bien huilée, calibrée pour prendre un grand nombre de décisions similaires, sans réelles contraintes de temps. L’analyse stratégique, en revanche, est un travail *ad hoc* sur un cas particulier, et la décision doit être prise rapidement. Plusieurs membres du comité avertirent Joan qu’ils s’opposeraient à toute procédure qui compliquerait exagérément l’analyse et retarderait la décision. Ils s’inquiétaient aussi d’ajouter une charge supplémentaire aux efforts déjà conséquents fournis par l’équipe d’analystes de Mapco.

Joan répondit directement à ces objections. Elle promit à ses collègues que structurer la décision ne la retarderait pas pour autant. « La date de la réunion du conseil où nous parlerons de l’opération reste la même ; tout ce que nous allons changer, c’est son ordre du jour, expliqua-t-elle. Nous allons commencer par dresser une liste des différents aspects de l’opération qu’il convient d’évaluer, de même qu’un recruteur commence par établir une description de poste pour avoir une liste de ce qu’on recherche chez un candidat. Nous nous assurerons ensuite que le conseil d’administration discute séparément de chacune de ces évaluations intermédiaires, exactement comme un recruteur, dans un entretien structuré, évalue successivement les différents aspects d’une candidature. Alors, et seulement alors, nous discuterons pour savoir si nous faisons l’opération ou pas. Cette procédure sera un moyen beaucoup plus efficace de tirer parti de l’intelligence collective du conseil d’administration.

» Si nous sommes d’accord sur ce principe, bien sûr, cela change la façon dont les informations devront être présentées et dont l’équipe chargée

de préparer l’opération devra travailler. C’est pourquoi je voulais avoir votre opinion dès maintenant. »

L’un des membres du comité n’était toujours pas convaincu. Il demanda à Joan en quoi la structuration améliorerait la qualité des décisions de recrutement, et pourquoi elle pensait que ces améliorations pourraient aussi s’appliquer à une décision stratégique. La question donna à Joan l’occasion de détailler sa logique. Le protocole d’évaluations intermédiaires, expliqua-t-elle, maximise la valeur des informations en garantissant que les différents aspects de l’évaluation restent indépendants les uns des autres. « Nos discussions habituelles en conseil ressemblent beaucoup à des entretiens non structurés, observa-t-elle. Nous avons toujours à l’esprit notre objectif final, qui est de prendre une décision, et nous traitons toutes les informations au regard de cet objectif. Nous cherchons à arriver à une conclusion, et dès qu’il s’en présente une, nous avons tendance à nous en satisfaire. Comme un recruteur dans un entretien non structuré, nous risquons de consacrer toute la discussion à confirmer nos premières impressions.

» L’intérêt de la structuration, c’est qu’elle nous force à repousser la prise de décision jusqu’à ce que nous ayons procédé à toutes les évaluations nécessaires. Ces évaluations séparées seront nos objectifs intermédiaires. De cette façon, nous sommes sûrs de prendre en compte toutes les informations disponibles. Et nous évitons que notre conclusion sur les premiers aspects de la transaction que nous discutons ne parasite notre lecture des autres aspects, même quand ils sont sans rapport. »

Les membres du comité acceptèrent l’expérimentation. Mais, demandèrent-ils aussitôt à Joan, en quoi consisteraient ces fameuses évaluations intermédiaires ? Joan avait-elle déjà pensé à une sorte de check-list compilant les critères de décision ? « Non, répondit-elle. Ce pourrait être le cas si nous parlions d’une décision de routine, à prendre de manière répétitive. Mais ici, c’est à nous de définir les critères intermédiaires. C’est

même l'étape la plus importante : c'est à nous qu'il appartient de déterminer sur quels éléments nous déciderons ou non de réaliser l'opération. » Le comité stratégique convint de se réunir à nouveau le lendemain pour ce faire.

## *Deuxième réunion : définir les critères d'évaluation intermédiaire*

« Notre première tâche, expliqua Joan, c'est de dresser une liste des critères d'évaluation possibles pour cette transaction, qu'il appartiendra à l'équipe d'analystes de Jeff Schneider d'évaluer. Notre travail aujourd'hui est de bâtir une liste exhaustive de critères indépendants. *Exhaustive,* au sens où tout fait qui vous semble pertinent pour la décision devrait influencer au moins un des critères. *Indépendants* signifie que, de préférence, il ne devrait pas y avoir de recouvrement entre les critères, afin d'éviter des évaluations redondantes. »

Le groupe se mit au travail et produisit rapidement une longue liste de faits et de données qui lui semblaient pertinents. Il les organisa ensuite pour en faire une liste de critères à évaluer. La liste devait être à la fois courte, complète et composée de critères d'évaluation qui ne se chevauchaient pas. L'exercice n'était pas facile, mais pour des dirigeants expérimentés, il était faisable. Au bout du compte, le groupe se mit d'accord sur une liste de sept critères d'évaluation, qui ressemblait un peu à la table des matières du rapport qu'un conseil d'administration, d'ordinaire, recevrait avant de discuter d'une acquisition. Outre la modélisation financière de rigueur, la liste des critères comprenait, par exemple, la qualité de l'équipe de direction

de l’entreprise cible et la probabilité que les synergies anticipées se réalisent.

Certains membres du comité furent déçus de constater que la réunion n’avait pas apporté de révélations sur Roadco. Mais ce n’était nullement son objectif, leur répondit Joan. Il ne s’agissait pas de trouver des réponses, mais de savoir quelles questions se poser ; et plus précisément, de briefer l’équipe projet chargée d’étudier le dossier. Chacun des sept critères d’évaluation intermédiaire ainsi définis, ajouta-t-elle, ferait l’objet d’un chapitre différent dans le rapport de l’équipe projet, et le conseil d’administration en débattrait séparément.

En d’autres termes, pour Joan, la mission de l’équipe projet n’était pas de donner au conseil une recommandation générale sur l’opportunité de l’acquisition – du moins, pas encore. Elle consistait à lui fournir une évaluation indépendante et objective *sur chacun des critères d’évaluation intermédiaire*. En fin de compte, expliqua Joan, l’équipe projet chargée de l’opération devrait conclure chaque chapitre de son rapport par une note quantifiant la réponse à une question simple : « Indépendamment du poids que ce critère doit avoir dans la décision finale, dans quelle mesure les éléments que nous avons recueillis sur ce critère d’évaluation militent-ils pour ou contre l’acquisition ? »

## *Le rôle de l’équipe projet d’évaluation*

Le chef de l’équipe projet chargée d’évaluer l’opération, Jeff Schneider, réunit son équipe dans l’après-midi pour organiser le travail. Les changements apportés à la manière habituelle de travailler n’étaient pas très nombreux, mais Jeff voulait s’assurer que son équipe y serait attentive.

Il souligna tout d’abord l’importance pour les analystes de son équipe de réaliser des analyses aussi objectives que possible. Non seulement les

évaluations devraient reposer sur des faits – rien de neuf ici – mais elles devraient aussi adopter une *vision externe* chaque fois que c'était possible. Comme les membres de l'équipe demandaient ce que signifiait la « vision externe », Jeff leur en donna deux exemples, à l'aide de deux des critères d'évaluation intermédiaire identifiés par Joan.

L'un des critères consistait à évaluer la probabilité que la transaction reçoive l'accord des autorités de la concurrence. Adopter la vision externe, expliqua Jeff, implique que l'équipe commence par rechercher le *taux de base*, c'est-à-dire le pourcentage de transactions comparables qui sont approuvées. Et pour ce faire, l'équipe devrait définir une *classe de référence*, comprenant toutes les transactions récentes jugées suffisamment comparables.

Jeff donna un second exemple de vision externe, en expliquant comment évaluer les compétences technologiques de la direction R&D de Roadco – un autre critère d'évaluation figurant sur la liste de Joan. « Il ne suffit pas d'énumérer les réussites récentes de l'entreprise et de les qualifier de "bonnes" ou de "remarquables". Ce que j'attends de vous, c'est une évaluation relative, quelque chose comme : "Par rapport aux entreprises comparables que nous connaissons, et après avoir évalué les résultats des lancements de produits des trois dernières années, notre jugement est que les compétences de cette équipe la placent dans le deuxième quintile". » De manière générale, expliqua Jeff, chaque fois que c'est possible, les évaluations devraient être comparatives, parce que les jugements relatifs sont plus fiables que les jugements absolus.

Jeff donna une autre instruction importante à son équipe. Joan, expliqua-t-il, insistait pour que les évaluations intermédiaires soient aussi indépendantes les unes des autres que possible, afin de réduire le risque qu'elles ne s'influencent mutuellement. C'est pourquoi Jeff donna à des membres différents de son équipe la responsabilité de différents critères sur la liste de Joan, en insistant pour qu'ils y travaillent séparément.

Certains analystes de l’équipe ne cachèrent pas leur surprise. « Pourquoi ne devrions-nous pas travailler en équipe ? demandèrent-ils à leur chef. À quoi bon réunir une équipe si tu ne veux pas que nous communiquions entre nous ? »

Jeff prit le temps d’expliquer l’importance de l’indépendance dans les jugements. « Vous connaissez tous l’effet de halo dans le recrutement, commença-t-il. C’est ce qui se passe quand l’impression générale que produit un candidat influence l’évaluation de tel ou tel aspect précis de ses compétences. C’est ce que nous voudrions éviter. » Mais certains analystes semblaient toujours penser que le problème n’était pas bien grave. Jeff recourut donc à une autre analogie. « Si vous enquêtiez sur un crime et s’il y avait quatre témoins, les laisseriez-vous discuter entre eux avant de les interroger ? Bien sûr que non ! Vous ne voudriez pas qu’un témoin influence les autres. » Les analystes ne trouvèrent pas la comparaison très flatteuse, mais ils comprirent le message.

En pratique, Jeff n’avait pas assez d’analystes dans son équipe pour que chaque critère d’évaluation soit traité de manière parfaitement indépendante. Jane, membre expérimenté de l’équipe, fut donc chargée de deux des évaluations intermédiaires. Jeff les choisit aussi différentes que possible et demanda à Jane de les traiter consécutivement, et de rédiger sa conclusion sur la première avant de passer à la seconde. Une autre préoccupation pratique était d’évaluer la qualité de l’équipe de direction de Roadco : Jeff craignait que ses analystes aient du mal à porter un jugement sur la qualité intrinsèque de l’équipe de direction, distinct de leur évaluation des résultats récents de l’entreprise (qu’ils devraient, bien sûr, étudier en détail). Pour remédier au problème, Jeff demanda à un expert en ressources humaines extérieur de participer à cette évaluation. De cette façon, pensait-il, les évaluations des différentes dimensions seraient plus indépendantes.

Une autre recommandation de Jeff parut inhabituelle à son équipe. Chaque chapitre devait porter sur un des critères d’évaluation intermédiaire

et, comme le souhaitait Joan, déboucher sur une conclusion, formulée sous forme de note. Mais, ajouta Jeff, les analystes devraient aussi inclure dans chacun de ces chapitres la totalité des informations factuelles pertinentes sur l'évaluation concernée. « Ne dissimulez rien, leur dit-il. Le ton général du chapitre sera cohérent avec la note proposée, bien sûr, mais s'il y a des informations qui semblent incohérentes ou même contradictoires avec la note principale, ne mettez rien sous le tapis. Votre mission n'est pas de vendre vos recommandations en construisant une belle histoire simple et claire. C'est de représenter la réalité. Si elle est compliquée et ambiguë – comme c'est souvent le cas – dites-le au conseil. Il a besoin de le savoir. »

Dans le même esprit, Jeff encouragea les analystes à faire preuve de transparence sur leur niveau de confiance dans chaque évaluation. « Le conseil d'administration sait que les informations dont vous disposez ne sont pas parfaites. Ce qu'il ne sait pas forcément, c'est sur quoi vous avez réussi à réunir beaucoup d'informations, et sur quoi vous êtes vraiment dans le noir. Cela les aidera beaucoup que vous le leur disiez. Et bien sûr, si vous tombez sur quelque chose qui pose un vrai problème – un élément susceptible de nous faire renoncer à la transaction –, vous le signalerez immédiatement. »

L'équipe chargée de l'opération procéda comme on le lui avait demandé. Elle ne trouva pas d'éléments susceptibles de faire capoter la transaction, et poursuivit donc ses évaluations, en rédigeant un rapport pour Joan et pour le conseil d'administration, qui couvrait tous les critères d'évaluation retenus.

## *Troisième réunion : la décision*

En lisant le rapport préparé par l'équipe de Jeff Schneider pour le conseil d'administration, Joan remarqua immédiatement quelque chose

d’important : si la plupart des évaluations intermédiaires plaidaient en faveur de l’acquisition, elles n’étaient pas uniformément positives. Certaines des notes étaient élevées, d’autres non. En somme, le rapport ne présentait pas de l’opération proposée une analyse univoque, invitant à une conclusion unique et à une action immédiate : le tableau était plus contrasté que d’ordinaire.

Joan n’en fut pas surprise. Quand les évaluations sont faites indépendamment les unes des autres, l’excès de cohérence est neutralisé. La réalité apparaît, avec sa complexité, ses incertitudes et ses ambiguïtés. Elle n’est pas aussi cohérente qu’il y paraît dans la plupart des présentations faites à des conseils d’administration.

« Tant mieux, se dit Joan. Ces disparités entre les évaluations intermédiaires vont soulever des questions et susciter des discussions. C’est exactement ce dont nous avons besoin pour avoir un bon débat au sein du conseil. Les contrastes ne rendront pas la décision plus facile, c’est sûr, mais ils la rendront meilleure. »

Joan réunit donc le conseil d’administration pour examiner le rapport et prendre une décision sur l’acquisition. Elle expliqua l’approche suivie jusque-là, et invita les membres du conseil à appliquer le même principe : « Jeff et son équipe ont fait beaucoup d’efforts pour que les évaluations restent indépendantes les unes des autres, dit-elle, et notre travail aujourd’hui est de les examiner, à notre tour, de manière tout aussi indépendante. Cela veut dire que nous allons discuter de l’évaluation intermédiaire sur chaque critère, l’un après l’autre. C’est seulement ensuite que nous discuterons de la décision globale. De fait, nous allons faire comme si chaque critère d’évaluation était un sujet distinct à l’ordre du jour de notre réunion. »

Les membres du conseil savaient qu’il ne serait pas facile de suivre cette approche structurée. Joan leur demandait de résister à la tentation de se former une opinion globale sur l’opération avant d’avoir discuté des

évaluations intermédiaires, mais nombre d'entre eux, familiers du secteur, avaient *déjà* une opinion sur Roadco. Ne pas s'y référer leur semblait un peu artificiel. Cependant, comme ils comprenaient ce qu'essayait de faire Joan, ils acceptèrent de jouer le jeu et de remettre à plus tard l'expression de leur point de vue global sur la transaction.

Les membres du conseil furent les premiers surpris de le constater : grâce à la méthode suggérée par Joan, la discussion fut beaucoup plus intéressante que d'habitude. Certains d'entre eux, arrivés avec une opinion qu'ils pensaient déjà faite, changèrent même d'avis sur la pertinence de l'opération (mais personne ne s'en rendit compte, puisqu'ils avaient gardé leurs opinions pour eux). La technique utilisée par Joan pour mener les débats joua un rôle considérable, notamment parce qu'elle eut recours à la méthode *estimation-discussion-estimation*[2], qui allie les avantages de la délibération et ceux de l'agrégation d'opinions indépendantes.

Voici comment elle procéda. Pour chacun des critères d'évaluation, Jeff Schneider commença par résumer brièvement les principaux faits (que les membres du conseil avaient préalablement lus en détail dans son rapport). Puis Joan demanda aux administrateurs de se servir d'une application sur leur smartphone pour donner leur propre note sur le critère : soit la même note que celle proposée par l'équipe de Jeff, soit une note différente. La distribution des notes fut aussitôt projetée sur l'écran (sans montrer qui avait donné chaque note). « Il ne s'agit pas encore d'un vote, précisa Joan. Nous prenons seulement la température du groupe sur chaque sujet. » Ainsi, Joan obtint une photographie instantanée des opinions des administrateurs, *avant* qu'ils n'entament la discussion, et donc avant qu'ils ne soient soumis aux risques d'influence sociale et de cascade informationnelle.

Sur certaines évaluations intermédiaires, le consensus fut immédiat, mais sur d'autres, le processus fit apparaître des opinions opposées. Naturellement, Joan fit en sorte que le conseil consacre plus de temps aux secondes. Elle s'assura que les membres du conseil puissent tous

s’exprimer, quel que fût leur avis, et les encouragea à donner leur opinion avec des faits et des arguments, mais aussi avec nuance et humilité. Et quand un des membres qui était fortement opposé à la transaction s’emporta, elle rappela au conseil : « Nous sommes tous des personnes raisonnables et il se trouve que nous ne sommes pas d’accord ; c’est donc qu’il doit s’agir d’un sujet sur lequel des gens raisonnables peuvent être en désaccord. »

Au terme de la discussion sur chaque critère d’évaluation intermédiaire, Joan demanda aux membres du conseil de voter une seconde fois. La plupart du temps, il y avait plus de convergence qu’au premier tour. La même séquence – une première estimation « à froid », une discussion, une seconde estimation – se répéta pour chaque critère d’évaluation.

Le moment vint enfin de conclure quant à l’opportunité de la transaction. Pour faciliter la discussion, Jeff afficha au tableau la liste des critères d’évaluation intermédiaire, et, pour chacun d’entre eux, la moyenne et la distribution des notes données par les membres du conseil. Le tableau présentait donc un « profil » synthétique de la décision à prendre. Mais comment les administrateurs allaient-ils prendre la décision ?

L’une des participantes fit une suggestion : prendre une simple moyenne arithmétique des notes. (Peut-être était-elle au fait de la supériorité de l’agrégation mécanique sur le jugement clinique global…) Un autre objecta aussitôt que, pour procéder de cette manière, il faudrait d’abord décider de la pondération des critères, et précisa lesquels, à son avis, devraient avoir davantage de poids que les autres. Une troisième administratrice approuva la démarche, mais en proposant une hiérarchisation différente des critères.

Joan interrompit la discussion. « Nous ne sommes pas là pour calculer une moyenne pondérée des évaluations, leur dit-elle. Nous avons différé le moment de recourir à nos intuitions, mais ce moment est venu. Ce dont nous avons besoin maintenant, c’est de votre jugement. »

Joan n’expliqua pas la logique qui sous-tendait cette suggestion, mais elle avait appris cette leçon à ses dépens. Elle savait que les dirigeants détestent tous les systèmes qui leur lient les mains et leur interdisent de faire usage de leur jugement – en particulier quand il s’agit de décisions importantes. Elle avait vu qu’il est facile pour les décideurs de contourner ce type de système : quand ils savent qu’une formule va être utilisée pour calculer mécaniquement la décision finale, il leur suffit de changer les notes qu’ils donnent à chaque critère pour arriver à la conclusion souhaitée – ce qui rend bien sûr tout l’exercice inutile.

En outre, même si ce n’était pas le cas ici, Joan restait attentive à la possibilité que surgissent des considérations décisives non prises en compte dans les critères d’évaluation intermédiaire (ce que nous avons appelé, au chapitre 10, des « jambes cassées »). Si des facteurs de blocage inattendus, ou, à l’inverse, des facteurs décisifs faisant pencher la balance en faveur de la transaction, étaient apparus en cours de route, un processus de décision mécanique, reposant sur la moyenne des évaluations intermédiaires, aurait pu déboucher sur une erreur grave.

Surtout, Joan savait qu’inviter les membres du conseil à faire usage de leur intuition à ce stade du processus n’était pas du tout la même chose que le faire plus tôt. Maintenant que les évaluations intermédiaires avaient été réalisées indépendamment les unes des autres et revues par chacun des membres du conseil, maintenant qu’elles étaient connues de tous et affichées, la décision finale de chacun serait nécessairement ancrée dessus. L’intuition serait fondée sur ces notes, reposant elles-mêmes sur des faits, et déjà discutées de manière approfondie. Face à une liste d’évaluations intermédiaires qui penchent dans un sens, un membre du conseil pourrait toujours pencher dans l’autre, mais il lui faudrait des arguments puissants. En suivant cette logique, le conseil s’engagea donc dans une dernière discussion sur la transaction, et vota, comme tous les conseils d’administration, sur la proposition qui lui était soumise.

## *Le protocole d'évaluations intermédiaires dans les décisions récurrentes*

Nous avons décrit le protocole d'évaluations intermédiaires dans le contexte d'une décision « singulière » : une décision stratégique qui soulève des problèmes uniques et qui n'est prise qu'une seule fois. Mais la procédure s'applique tout aussi bien aux décisions récurrentes. Imaginons que Mapco ne soit pas une entreprise industrielle qui réalise une acquisition exceptionnelle, mais un fonds de capital-risque qui investit régulièrement dans des start-up. Le protocole serait tout aussi applicable et fonctionnerait à peu près de la même manière, à deux détails près qui, d'ailleurs, le rendent plus facile à appliquer.

D'abord, la première étape – la définition de la liste des critères d'évaluation intermédiaire – n'a besoin d'être effectuée qu'une seule fois. Le fonds a ses critères de décision : il les applique à tous les investissements considérés, et il est donc inutile de les réinventer chaque fois.

Ensuite, si le fonds évalue un grand nombre de sociétés et prend un grand nombre de décisions, il peut se servir de son expérience pour mieux calibrer ses jugements. Prenons l'exemple d'un critère essentiel pour tout fonds de capital-risque : l'évaluation de la qualité d'une équipe d'entrepreneurs. Bien sûr, évaluer la qualité d'une équipe n'est jamais chose facile, surtout si on n'a pas beaucoup d'occasions de pratiquer ce genre d'évaluation. Souvenez-vous, par exemple, que Jeff Schneider demandait à son équipe d'analystes d'évaluer la qualité du management de Roadco par rapport à une classe de référence : en théorie, la logique est inattaquable, mais en pratique, il n'est pas facile de recueillir des données sur des entreprises comparables.

Ce type de jugements comparatifs devient bien plus facile quand la décision est récurrente. Si le fonds de capital-risque a évalué les équipes de

direction de dizaines ou même de centaines de start-up, cette base d'expérience partagée devient sa classe de référence. Pour formaliser des jugements relatifs par rapport à cette classe de référence, le fonds peut créer une échelle de cas définie par des cas-ancres. Il pourrait, par exemple, évaluer les fondateurs de la start-up qu'il considère aujourd'hui comme « aussi performants que les deux fondateurs de la start-up ABC quand nous y avons investi », mais pas « aussi performants que la fondatrice de DEF ». Les cas-ancres (ici, ABC et DEF) doivent bien sûr être connus de tous les participants, et mis à jour régulièrement ; et leur formalisation nécessite un investissement en temps. Mais les jugements relatifs (comparer une équipe à celles d'ABC et DEF) seront beaucoup plus fiables que des notes absolues sur une échelle définie par des nombres ou des adjectifs.

## *Ce que change le protocole*

Par commodité, nous résumons dans le tableau 4 les principaux changements apportés par le protocole d'évaluations intermédiaires.

1. Au début du processus, structurer la décision en critères d'évaluation intermédiaire. *(Pour les jugements récurrents, on établit cette liste de critères une fois pour toutes.)*
2. Chaque fois que c'est possible, calibrer les évaluations intermédiaires par rapport à une vision externe. *(Pour les jugements récurrents, utiliser des jugements relatifs, si possible par rapport à une échelle de cas.)*
3. Dans la phase de collecte et d'analyse des informations, faire en sorte que les évaluations des différents critères restent aussi indépendantes les unes des autres que possible.
4. Pendant la réunion de décision, examiner chaque critère d'évaluation successivement et séparément.
5. Sur chaque critère, s'assurer que les participants produisent d'abord leurs jugements individuellement ; puis utiliser la méthode estimation-discussion-estimation.
6. Pour prendre la décision finale, différer le recours à l'intuition, mais ne pas l'interdire.

**Tableau 4.** Les principales étapes du protocole d'évaluations intermédiaires.

Vous aurez reconnu dans cette description une série de techniques d'hygiène de la décision présentées dans les précédents chapitres : séquencer l'information, structurer la décision en la décomposant en évaluations indépendantes, utiliser un cadre de référence commun reposant sur une vision externe, agréger les jugements indépendants de multiples individus. En formalisant l'application de ces techniques, le protocole d'évaluations intermédiaires vise à modifier le processus de décision pour y introduire la meilleure hygiène de la décision possible.

Nul doute que cet accent mis sur le *processus* de décision – plutôt que sur le contenu de la décision elle-même – fera froncer quelques sourcils. Les questions et les réactions que nous avons prêtées ici aux administrateurs et aux membres de l'équipe projet ne sont pas inhabituelles. Parler des tenants et des aboutissants d'une décision particulière, de ce qui la rend unique, est toujours plus intéressant que se pencher sur un processus de décision, nécessairement générique et passe-partout. Pour prendre une décision importante, utiliser son intuition et son jugement est valorisant ; suivre une méthode prédéfinie l'est moins. Par ailleurs, une idée reçue veut que les grandes décisions – et en particulier les décisions réellement

géniales – naissent de l'intuition et de la créativité des grands chefs. (Vous souscrirez sans doute plus volontiers à cette idée si vous êtes le grand chef en question.) Enfin, aux yeux de beaucoup, le mot *processus* évoque surtout la bureaucratie, les formulaires paperassiers et inutiles, et les décisions qu'on remet éternellement à plus tard.

Notre expérience des entreprises et des agences gouvernementales qui ont mis en place certains des éléments du protocole suggère que ces craintes sont injustifiées. Bien sûr, si une organisation est déjà paralysée par la bureaucratie, ajouter de la complexité à son processus de décision n'est pas la priorité. Mais il n'y a aucune fatalité à ce que l'hygiène de la décision ralentisse les décisions, et encore moins à ce qu'elle les bureaucratise. Au contraire, comme on l'a vu, l'un de ses principes essentiels est de susciter des jugements indépendants, et de provoquer le débat et la confrontation sur ces jugements. On est loin de la recherche étouffante de consensus et de l'évitement des conflits qui caractérisent les bureaucraties.

---

Le protocole d'évaluations intermédiaires est un exemple d'application d'hygiène de la décision. Certains l'utiliseront dans son entièreté ; d'autres se contenteront d'une ou deux techniques d'hygiène de la décision. Mais l'objectif et l'enjeu sont les mêmes.

En règle générale, les dirigeants d'entreprise et les responsables publics ignorent qu'il y a du bruit dans leurs décisions les plus importantes ; aussi ne font-ils rien pour le réduire. À cet égard, ils sont comparables aux recruteurs qui continuent à se reposer exclusivement sur des entretiens non structurés : sourds au bruit qui affecte leur jugement, exagérément confiants dans la qualité de leurs décisions, et ignorants des procédures qui pourraient les améliorer.

Comme le lavage des mains ne prévient pas toutes les maladies, l'hygiène de la décision n'empêchera jamais toutes les erreurs. Elle ne rendra certainement pas toutes les décisions brillantes. Mais comme le

lavage des mains, l'hygiène de la décision répond à un problème grave, invisible et pourtant omniprésent. Dès lors qu'il y a jugement, il y a bruit – et l'hygiène de la décision peut le réduire.

### À propos du protocole d'évaluations intermédiaires

« Nous avons un protocole structuré pour prendre des décisions de recrutement. Pourquoi ne pas en avoir un pour les décisions stratégiques ? Après tout, les options ne sont-elles pas comme les candidats ? »

« C'est une décision difficile. Quels sont les critères d'évaluation intermédiaire sur lesquels elle devrait reposer ? »

« Notre jugement intuitif global sur ce projet est important – mais n'en parlons pas pour l'instant. Notre intuition nous sera plus utile une fois qu'elle sera informée par les évaluations séparées que nous avons demandées. »

PARTIE VI

# Le bruit optimal

En 1973, le juge Marvin Frankel appelait à des actions énergiques pour réduire le bruit dans la fixation des peines criminelles. Il avait raison. Son audit de bruit informel avait mis en lumière des disparités injustifiées dans le traitement d'individus similaires, un résultat que des études plus systématiques devaient confirmer et amplifier. Ces disparités étaient scandaleuses. Et de plus, elles étaient frappantes pour ceux qui les découvraient.

L'un des objectifs de ce livre, on l'a vu, est de généraliser les arguments du juge Frankel et de proposer une analyse des fondements psychologiques du bruit. Beaucoup trouveront sans doute que le bruit dans la justice pénale est particulièrement intolérable. Mais il n'est pas tolérable non plus dans bien d'autres contextes, où des individus présumés interchangeables produisent dans l'exercice de leur métier des jugements radicalement différents. Dans l'assurance, le recrutement et l'évaluation des employés, la médecine, l'expertise médico-légale, l'enseignement, l'entreprise et l'administration, le bruit interpersonnel est une source d'erreur majeure. Nous avons vu aussi que nous sommes tous sujets au bruit occasionnel, c'est-à-dire que des facteurs qui ne devraient pas influer sur nos jugements

peuvent nous conduire à prendre des décisions différentes le matin et l'après-midi, ou le lundi et le jeudi.

Pourtant, comme le montre la réaction très négative du monde judiciaire américain à l'introduction de directives encadrant la fixation des peines, les tentatives pour réduire le bruit soulèvent souvent des objections sérieuses, voire passionnées. Dès qu'on parle de directives, beaucoup de gens les trouvent rigides, déshumanisantes et, en fin de compte, injustes. Il nous est tous arrivé de faire une demande raisonnable à une entreprise, à un employeur ou à une administration, et de nous voir opposer une réponse du type : « Nous aimerions vraiment vous aider, mais nous avons les mains liées, nous avons des règles à appliquer. » Les règles en question semblent souvent stupides et inhumaines, même si elles partent d'une bonne intention : réduire le bruit (et souvent aussi le biais).

Les tentatives de réduction du bruit sont particulièrement mal accueillies quand elles rendent difficile ou impossible pour un individu d'être entendu et de présenter sa position. La crainte d'être privé de cette possibilité, qui fait partie des garanties d'un procès équitable, est pour beaucoup dans les inquiétudes et les résistances que suscitent les algorithmes et l'intelligence artificielle. Personne ne défile sous des banderoles clamant : « Nous voulons des algorithmes ! »

Kate Stith, de la Yale Law School, et José Cabranes, un juge fédéral, ont pris à ce sujet une position qui a eu une certaine influence aux États-Unis. Leur critique féroce des directives en matière de peine criminelle peut se lire comme une charge contre l'un des principaux arguments que nous développons dans ce livre, car leurs arguments s'appliquent tout aussi bien à la réduction du bruit dans l'enseignement, dans l'entreprise, dans le sport, et partout ailleurs. Stith et Cabranes soutiennent que les directives en matière de peine sont inspirées par « la peur d'exercer un pouvoir discrétionnaire – la peur de juger – et par une foi technocratique dans l'expertise et la planification centrale ». Ils estiment que cette « peur de

juger » a pour effet d'interdire la prise en considération des « particularités de chaque affaire ». À leurs yeux, « aucune solution mécanique ne peut satisfaire les exigences de la justice[1] ».

Ces objections méritent d'être examinées. Dans beaucoup de contextes où des jugements sont nécessaires, on considère souvent les « exigences de la justice » comme incompatibles avec toute espèce de solution mécanique. Cette position revient à permettre, et même à imposer, des méthodes qui produiront inévitablement du bruit. Les appels à être attentif aux « particularités de chaque affaire » exercent un attrait puissant non seulement dans les tribunaux, mais aussi dans les hôpitaux, les écoles ou les entreprises. Et même quand on choisit des techniques d'hygiène de la décision qui ne consistent pas à imposer des solutions mécaniques – par exemple, quand on décompose un problème en plusieurs éléments, ou quand on agrège les jugements indépendants de plusieurs individus –, beaucoup de gens sont encore hostiles à l'utilisation de ces techniques.

Nous avons défini le bruit comme de la variabilité indésirable, et qualifier quelque chose d'indésirable suggère qu'il faut l'éliminer. Mais l'analyse est plus compliquée et plus intéressante que cela. Le bruit est sans doute indésirable, toutes choses égales par ailleurs ; mais toutes choses ne sont pas toujours égales par ailleurs. Les coûts de l'élimination du bruit peuvent être supérieurs à ses bénéfices. Qui plus est, même quand le bruit est extrêmement coûteux, son élimination peut avoir des effets secondaires importants.

Dans cette partie du livre, nous allons nous pencher sur les objections que soulèvent les efforts de réduction du bruit. Ces objections peuvent être regroupées en sept grandes catégories.

Premièrement : *réduire le bruit peut être trop coûteux*. Les gains qu'on en retire peuvent tout simplement ne pas en valoir la peine. Inversement, les mesures nécessaires pour réduire le bruit peuvent être extrêmement contraignantes, ou même inapplicables en pratique.

Deuxièmement : *certaines stratégies de réduction du bruit peuvent être elles-mêmes génératrices d'erreur*. En particulier, elles peuvent parfois produire du biais systémique. Si tous les prévisionnistes d'un ministère faisaient exactement la même prévision exagérément optimiste, il n'y aurait pas de bruit, mais les prévisions seraient toutes fausses. Ou, pour prendre un exemple plus évident encore, si tous les médecins d'un hôpital prescrivaient de l'aspirine à tous les patients, quels que soient leurs symptômes, ils ne seraient pas sujets au bruit – mais ils feraient évidemment beaucoup d'erreurs.

Nous répondrons à ces deux objections au chapitre 26. Dans le suivant, nous en examinerons cinq autres, plus complexes et elles aussi fréquentes. Ce sont des objections que nous allons sans doute entendre beaucoup dans les années qui viennent, à mesure que l'on recourra de plus en plus aux règles, aux algorithmes et à l'apprentissage automatique.

Troisièmement : *traiter les personnes avec respect et dignité impose de tolérer une certaine quantité de bruit.* Le bruit peut être un sous-produit d'un processus certes imparfait, mais qui donne à chacun la possibilité de plaider sa cause en tant qu'individu. C'est au prix du bruit que les employés, les clients, les candidats à un emploi, les étudiants ou les prévenus conservent la possibilité d'influencer le jugement discrétionnaire dont ils font l'objet, et le sentiment d'avoir été dûment entendus.

Quatrièmement : *le bruit peut être une condition de l'adaptation.* Tout système de jugement doit pouvoir s'adapter à des valeurs morales et politiques en perpétuelle évolution. En ce sens, éliminer le bruit pourrait affaiblir sa capacité à réagir à des situations nouvelles et inattendues. Un système exempt de bruit risquerait de figer les valeurs existantes.

Cinquièmement : *certaines stratégies de réduction du bruit pourraient favoriser des comportements opportunistes.* La clarté des règles pourrait paradoxalement être exploitée par certains pour contourner le système ou

éviter des interdits. Un peu de bruit pourrait donc être nécessaire pour empêcher certains actes répréhensibles.

Sixièmement : *le bruit pourrait être un bon outil de dissuasion*. Si vous êtes tenté par un comportement répréhensible, l'incertitude sur la sanction – modeste ou terrible – pourrait être en elle-même dissuasive, du moins si vous avez une certaine aversion au risque. En ce sens, un système qui intègre un peu de bruit pourrait être plus dissuasif.

Septièmement : *la réduction du bruit peut être source de démotivation.* Personne ne veut renoncer à la responsabilité d'exercer son jugement et se voir réduit au rôle de rouage dans une machine. Certaines stratégies de réduction du bruit, si elles donnent ce sentiment, peuvent saper l'initiative et la motivation des individus.

Nous allons bien sûr nous efforcer de répondre à ces objections avec autant d'objectivité et de compréhension que possible. Mais cela ne veut pas dire que nous les fassions nôtres. À tout le moins, nous ne pensons pas que ces objections suffisent à invalider l'objectif général de réduction du bruit.

La réponse que nous apporterons à ces objections, tout au long de cette dernière partie, est pour l'essentiel la suivante : *une objection à une stratégie de réduction du bruit ne vaut pas justification du bruit*. On peut avoir des raisons très convaincantes de s'opposer, dans un contexte particulier, à une technique particulière d'hygiène de la décision, sans pour autant en exclure d'autres. Par exemple, on peut reprocher leur rigidité aux directives tout en convenant que l'agrégation de jugements indépendants est une bonne idée. Ou trouver trop contraignant d'utiliser le protocole d'évaluations intermédiaires, tout en étant très favorable à la codification d'une échelle commune ancrée sur la vision externe.

Ayant tout cela à l'esprit, notre conclusion générale est que même quand ces objections sont examinées avec toute l'attention qu'elles méritent, la réduction du bruit reste un objectif important et urgent. Dans le chapitre 28,

nous défendrons cette conclusion en nous penchant sur un dilemme que nous rencontrons tous les jours, même si nous n’en sommes pas toujours conscients.

CHAPITRE 26

# Les coûts de la réduction du bruit

Suggérez à des professionnels qu'il serait bon d'éliminer le bruit : bien souvent, ils vous répondront que le jeu n'en vaut pas la chandelle. Parfois, ils soutiendront même que réduire le bruit n'est tout simplement pas possible. Nous avons entendu cette objection dans les entreprises, dans la sphère gouvernementale, dans l'enseignement et ailleurs. La préoccupation est légitime, mais il convient de ne pas l'exagérer. En vérité, elle n'est souvent qu'un alibi.

Pour présenter cette objection sous son jour le plus favorable, prenons le cas d'un professeur de lycée qui note chaque semaine vingt-cinq dissertations d'élèves de seconde. Si le professeur consacre une quinzaine de minutes à chaque copie, les notes qu'il leur attribuera seront sujettes au bruit, et donc sources d'injustice. Comment le professeur pourrait-il introduire un peu d'hygiène dans ses décisions de notation ? Il pourrait, par exemple, demander à un collègue de noter indépendamment les copies de ses élèves et utiliser la moyenne des deux corrections. Il pourrait aussi introduire plus de structuration dans son processus d'évaluation en définissant précisément son barème de notation (une forme de directive, ou de check-list). Il pourrait lire chaque copie deux fois, en s'assurant de les lire dans un ordre différent et en évitant bien sûr de faire figurer sur les

copies sa première note. Ou peut-être pourrait-il s'assurer qu'il lit toutes les copies à la même heure de la journée, pour réduire le bruit occasionnel.

Si nous nous mettons à la place de ce professeur, toutes ces précautions doivent lui sembler très excessives. À tort, ou à raison ? La réponse dépend de plusieurs facteurs. Y a-t-il beaucoup de bruit dans les notations du professeur ? Quel degré d'exactitude dans la notation peut-on espérer gagner ? Quel est l'enjeu de ce degré d'exactitude supplémentaire ? Combien de temps et d'argent seront nécessaires pour réduire le bruit ? Des réponses à ces questions dépend l'arbitrage à faire entre bénéfices et coûts de la réduction du bruit. On peut évidemment imaginer que cet arbitrage sera différent si l'on corrige des copies de concours plutôt que des dissertations de seconde. On peut même supposer qu'il ne sera pas le même dans toutes les matières, si l'on suppose que les notations sont moins sujettes au bruit dans certaines disciplines que dans d'autres.

Le principe de cette analyse peut être étendu aux situations plus complexes que rencontrent toutes sortes d'organisations publiques et privées. Ce raisonnement les conduira souvent à rejeter certaines stratégies de réduction du bruit, trop coûteuses au regard des enjeux. Dans le cas des diagnostics médicaux divergents, par exemple, l'enjeu de la réduction du bruit est évidemment important : faire moins d'erreurs peut sauver des vies. Mais le coût et surtout la faisabilité des efforts de réduction du bruit doivent, eux aussi, être pris en considération. Élaborer des directives, par exemple, n'est pas possible pour toutes les situations. Et même s'il existe des tests qui permettent de réduire le bruit dans les diagnostics, encore faut-il en mesurer les inconvénients : quand un test est invasif, risqué ou coûteux, et quand la variabilité des diagnostics n'a pas de conséquences graves, le test n'est sans doute pas justifié pour tous les patients.

L'évaluation des employés est rarement une question de vie ou de mort. Mais le bruit peut produire de l'injustice pour les employés et des coûts élevés pour l'entreprise. Nous avons vu qu'il existe des stratégies de

réduction du bruit qui sont techniquement réalisables. Beaucoup d'organisations pourraient toutefois penser qu'elles n'en valent pas la peine : éviter quelques évaluations erronées et démotivantes, pensent-elles, ne justifie pas une réforme élaborée du système d'évaluation, coûteuse à concevoir et pénible à mettre en œuvre. Il est clair que, dans certains cas, cette conclusion à courte vue est paresseuse et malavisée, et que l'organisation ferait bien d'envisager certaines formes d'hygiène de la décision. Cependant, le cas contraire existe aussi : l'idée qu'il est trop coûteux de réduire le bruit n'est pas toujours fausse.

En somme, la réduction du bruit a un coût ; et ses bénéfices attendus doivent lui être supérieurs. Cette objection est tout à fait sensée, et c'est une des raisons pour lesquelles les audits de bruit sont si importants. Dans un grand nombre de situations, ils permettent de révéler que le bruit produit des niveaux d'injustice inacceptables, des erreurs très coûteuses, ou les deux. Quand c'est le cas, les efforts nécessaires pour réduire le bruit ne justifient pas qu'on renonce à s'y attaquer.

## *Moins de bruit, plus d'erreurs ?*

Une autre objection est que les efforts de réduction du bruit peuvent eux-mêmes produire des niveaux d'erreur inacceptables. C'est certainement le cas quand les instruments utilisés pour combattre le bruit sont trop grossiers. Certaines mesures de réduction du bruit pourraient ainsi accroître le biais. Si, par exemple, un réseau social comme Facebook ou Twitter adopte des directives censurant tous les messages qui contiennent certains mots prédéfinis, ses décisions seront plus prévisibles et moins sujettes au bruit, mais beaucoup de messages seront retirés à tort (des « faux positifs »). On constatera alors, dans la politique de modération du réseau

social, beaucoup plus d'erreurs dans un sens que dans l'autre, c'est-à-dire un biais.

Un grand nombre de réformes institutionnelles visent à réduire le pouvoir d'appréciation discrétionnaire des individus et le bruit qui en résulte. Même quand ces réformes partent de bonnes intentions, il arrive que le remède soit pire que le mal. Dans *Deux siècles de rhétorique réactionnaire*, l'économiste Albert Hirschman souligne trois objections courantes à toute tentative de réforme[1]. D'abord, une réforme peut échouer et ne rien changer du tout. Ensuite, elle peut avoir des effets pervers, c'est-à-dire aggraver le problème qu'elle vise à résoudre. Enfin, elle peut mettre en péril d'autres valeurs importantes (par exemple quand on dit qu'une mesure visant à protéger la liberté syndicale nuit à la croissance économique). Ces trois objections peuvent être opposées à la réduction du bruit. La critique est quelquefois purement rhétorique, et ne vise qu'à faire capoter une réforme qui aurait en réalité beaucoup de bienfaits. Mais dans certains cas, les dangers sont réels, au sens où certaines stratégies de réduction du bruit pourraient mettre en danger des valeurs importantes ou avoir de sérieux effets pervers.

Ce sont des objections de ce type que formulaient les juges américains quand ils protestaient contre les directives encadrant la fixation des peines. Ils connaissaient le travail du juge Frankel et ne niaient pas que l'exercice de leur pouvoir d'appréciation discrétionnaire fût une source de bruit. Ils pensaient cependant qu'y porter atteinte ne produirait pas moins mais plus d'erreurs. Citant Václav Havel, ils affirmaient : « Nous devons renoncer à l'idée arrogante que le monde n'est qu'un problème à résoudre, une machine dont le mode d'emploi attend d'être découvert, un ensemble d'informations à rentrer dans un ordinateur dans l'espoir qu'il trouve un jour ou l'autre une solution universelle[2]. » Derrière ce rejet des « solutions universelles » se devine une idée largement répandue : parce que les situations humaines sont d'une infinie variété, seuls de bons juges, dans leur

sagesse, sont capables d'y faire face. Cela implique nécessairement de tolérer le bruit ou, tout au moins, de rejeter certaines des stratégies qui viseraient à le réduire.

Il est incontestable que la réduction du bruit peut entraîner des erreurs. À l'époque où les logiciels d'échecs étaient encore balbutiants, le système de divertissement en vol d'une grande compagnie aérienne proposait aux passagers de ses vols internationaux de jouer aux échecs contre un ordinateur. Au niveau de jeu le plus bas, le programme utilisait une règle simple : faire échec au roi de l'adversaire chaque fois que possible. Le programme n'était pas sujet au bruit. Il jouait tout le temps de la même manière en suivant toujours sa règle. Mais cette règle lui faisait commettre beaucoup d'erreurs. Le programme jouait très mal aux échecs, si mal que même des joueurs médiocres pouvaient le battre (c'était même sans doute l'objet de la chose : des passagers qui gagnent sont des passagers heureux).

Un exemple autrement plus lourd de conséquences est celui de la politique pénale adoptée dans certains États des États-Unis et appelée « loi des trois prises » ou « de la troisième infraction » (*Three strikes and you're out* [3]). L'idée est qu'un individu reconnu coupable pour la troisième fois d'un crime ou d'un délit soit obligatoirement condamné à la prison à vie – sans exception. Les partisans des lois du type « trois prises » voulaient réduire la variabilité entre juges, notamment parce qu'ils s'inquiétaient du bruit de niveau : certains juges, à leurs yeux, étaient systématiquement laxistes avec des délinquants endurcis. En d'autres termes, ces lois ont pour objectif central d'éliminer le bruit ; et si elles sont appliquées, évidemment, elles y parviennent.

Mais, même si l'objectif de réduction du bruit est atteint, il est permis de penser que le prix de ce succès est bien trop élevé. Certains individus qui ont commis trois délits ne méritent pas la perpétuité. Peut-être leurs délits ou leurs crimes n'étaient-ils pas violents. Peut-être devraient-ils bénéficier de circonstances atténuantes. Peut-être faudrait-il prendre en compte les

situations de vie qui ont contribué à les faire déraper. Peut-être, enfin, ces délinquants sont-ils capables de se réhabiliter. Pour toutes ces raisons, et d'autres, beaucoup de juristes et de citoyens considèrent qu'imposer une peine de prison à vie sans tenir compte des circonstances est non seulement trop sévère, mais surtout d'une rigidité intolérable. C'est pourquoi le prix de cette stratégie de réduction du bruit est trop élevé.

Autre exemple : le célèbre cas *Woodson vs North Carolina*[4]. Dans cette affaire, la Cour suprême des États-Unis a décidé qu'une loi imposant l'application obligatoire de la peine de mort dans certaines circonstances était inconstitutionnelle, non pas parce qu'elle était trop sévère, mais *parce que c'était une règle*. Toute l'idée de la loi en question était de réduire le bruit : dans certains cas précis, le meurtrier devait obligatoirement être condamné à mort. La Cour suprême invalida cette loi en invoquant la nécessité d'un traitement individualisé. « La croyance ne prévaut plus, écrivit la Cour, que toutes les infractions pour lesquelles une même qualification pénale est retenue nécessitent une sanction identique sans que soient prises en compte la vie passée et la personnalité de l'auteur des faits. » Selon la Cour suprême, un des motifs d'inconstitutionnalité de la peine de mort obligatoire est qu'elle « ne traite pas les personnes convaincues d'un crime ou d'un délit comme des êtres humains et des individus uniques, mais comme les membres d'une masse indifférenciée et sans visage qui doivent être condamnés de manière aveugle à la peine de mort ».

La peine de mort est bien sûr un sujet d'une gravité extrême, mais l'analyse de la Cour suprême des États-Unis peut être appliquée à beaucoup d'autres situations, même quand elles n'ont aucun rapport avec le droit. Les professeurs qui notent des élèves, les médecins qui diagnostiquent des patients, les employeurs qui évaluent des salariés, les assureurs qui fixent des primes d'assurance, les entraîneurs qui sélectionnent des sportifs : tous ces professionnels peuvent faire des erreurs s'ils tentent de réduire le bruit

en appliquant des règles d'une rigidité excessive. Si par exemple un employeur utilise des règles trop simplistes pour décider qui promouvoir ou qui licencier, il éliminera le bruit, mais négligera des aspects importants de la performance de ses collaborateurs. Il ne fait aucun doute qu'un système d'évaluation exempt de bruit mais incapable de prendre en compte des variables importantes peut être pire encore qu'un système reposant sur des jugements individuels sujets au bruit.

Nous reviendrons au chapitre 27 sur la distinction entre « individus uniques » et « membres d'une masse indifférenciée et sans visage ». Nous allons pour le moment nous pencher sur une question plus prosaïque : celle des stratégies de réduction du bruit qui, comme notre mauvais programme d'échecs, produisent trop d'erreurs.

Bien que nous venions d'en donner plusieurs exemples, l'objection n'est en réalité pas aussi convaincante qu'il n'y paraît. Car le fait qu'une stratégie de réduction du bruit entraîne des erreurs ne signifie pas qu'il faille se résoudre à des niveaux de bruit élevés. La solution n'est pas de renoncer à réduire le bruit, mais d'élaborer une meilleure stratégie de réduction du bruit. On peut par exemple envisager d'agréger plusieurs jugements humains plutôt que d'adopter des règles, ou choisir d'élaborer des directives avisées plutôt que des règles simplistes ou des solutions toutes faites. (Et l'on peut développer un meilleur programme d'échecs !)

Une université, par exemple, commence souvent, pour réduire le bruit dans ses décisions d'admission, par retenir les candidats ayant les notes les plus élevées. Mais si la règle paraît trop grossière, l'université peut créer une formule qui prend en compte non seulement les notes aux examens, mais aussi les résultats scolaires passés, les résultats sportifs, la situation familiale, et bien d'autres facteurs encore. Des règles complexes peuvent être plus appropriées, si elles prennent en compte une plus grande variété de facteurs considérés comme pertinents. De même, le diagnostic de certaines

maladies exige des médecins qu'ils suivent des règles bien plus complexes que celles du score d'Apgar que nous avons décrit.

En somme, les directives et les règles utilisées par les professionnels ne sont pas toujours grossières, et même pas nécessairement simples. C'est pourquoi un grand nombre d'entre elles contribuent à réduire le bruit sans se traduire par des coûts (ou des biais) intolérables. Et si les directives et les règles ne suffisent pas à produire de bons résultats, peut-être faut-il adopter d'autres techniques d'hygiène de la décision plus adaptées à la situation : par exemple, agréger les jugements indépendants de plusieurs personnes, ou utiliser un processus structuré comme le protocole d'évaluations intermédiaires.

## *Des algorithmes silencieux mais biaisés*

Quand on parle des méthodes de réduction du bruit que l'on soupçonne de produire des effets pervers, la question des algorithmes n'est jamais loin. En effet, les objections aux « biais algorithmiques » sont de plus en plus nombreuses. Comme on l'a vu, les algorithmes ont l'avantage d'éliminer le bruit ; et en ce sens, une partie de ce livre peut être considérée comme un plaidoyer pour un plus grand recours aux algorithmes. Cependant, l'utilisation d'un algorithme peut avoir des effets pervers intolérables si la réduction du bruit que permet l'algorithme est obtenue au prix d'une aggravation des discriminations sur la base de la race, du genre, ou de l'appartenance à des groupes défavorisés.

Cette crainte est répandue, et le risque que les algorithmes aient des effets discriminatoires est en effet incontestable. Dans *Weapons of Math Destruction* (« Armes de destruction matheuses »), la mathématicienne Cathy O'Neil sonne l'alarme contre le recours aux big data et aux algorithmes qui les exploitent, qu'elle accuse d'enraciner les préjugés,

d’accroître les inégalités et de menacer la démocratie elle-même[5]. Un autre sceptique note que « des modèles mathématiques potentiellement biaisés sont en train de remodeler nos vies, et ni les entreprises qui les développent ni le gouvernement n’ont envie de s’attaquer au problème[6] ». Autre exemple largement cité : selon ProPublica, une organisation indépendante de journalistes, un algorithme nommé Compas, largement utilisé aux États-Unis pour assister les juges dans l’évaluation du risque de récidive, est fortement biaisé contre les membres des minorités raciales[7].

Personne ne devrait en douter : il est possible – et même très facile – de créer un algorithme exempt de bruit, mais raciste, sexiste, ou sujet à un autre biais. Un algorithme qui utilise explicitement la couleur de peau d’un prévenu pour le mettre ou non en liberté sous caution est discriminatoire (et serait bien sûr illégal dans de nombreux pays). Un algorithme qui prend en compte la probabilité que des candidates à l’embauche demandent à bénéficier d’un congé maternité est évidemment discriminatoire aussi. Dans de tels cas et dans bien d’autres, des algorithmes pourraient certes éliminer la variabilité indésirable des jugements, mais y introduiraient des biais plus indésirables encore.

En principe, il devrait être possible de concevoir un algorithme qui ne prenne pas en compte la race ou le genre. On pourrait même exiger qu’il n’en tienne absolument pas compte. Mais la plus grande difficulté, qui fait aujourd’hui l’objet d’une attention particulière, est qu’un algorithme peut s’avérer discriminatoire et biaisé, même quand il *n’utilise pas* la race ni le genre comme sources d’information.

Comme nous l’avons noté plus haut, ce problème peut survenir pour deux raisons majeures. D’abord, l’algorithme peut, intentionnellement ou non, utiliser des prédicteurs apparemment neutres, mais en réalité fortement corrélés avec la race ou le genre. Par exemple, la taille et le poids sont corrélés avec le genre, et l’endroit où grandissent et vivent les gens est souvent corrélé avec la race.

Mais la discrimination peut aussi provenir de l'origine des données. Si un algorithme est entraîné sur une base de données biaisée, il sera nécessairement biaisé à son tour. Prenons les algorithmes de « police prédictive[8] », qui s'efforcent de prédire où et quand se produiront des actes de délinquance, souvent dans le but d'optimiser l'allocation des ressources de la police. Si les données existantes sur la délinquance reflètent une présence policière disproportionnée dans certains quartiers ou une surreprésentation de certaines catégories de délits dans les statistiques policières, alors les algorithmes qui reposent sur ces données risquent fort de perpétuer, voire d'exacerber, les discriminations. Il s'ensuit que même un algorithme qui ne prend pas expressément en compte la race ou le genre peut s'avérer aussi biaisé que des êtres humains, parce qu'il « apprend » à partir des mêmes données. D'une certaine façon, le résultat produit par un algorithme pourrait même sembler pire : quand on élimine le bruit, les biais sont plus visibles. Les algorithmes pourraient, en somme, être plus sûrement et plus constamment biaisés que des juges humains.

Pour beaucoup d'observateurs, une des principales questions pratiques que soulèvent les algorithmes est de savoir s'ils produisent sur certains groupes identifiables des effets différents (*disparate impact*). On parle parfois à ce sujet de discrimination non intentionnelle, par opposition à la discrimination intentionnelle (*disparate treatment*), qui applique à ces groupes des critères différents. On pourrait penser qu'il est facile de définir ce qui constitue un « effet différent », et donc ce qui rend un algorithme *de facto* discriminatoire, ou au contraire équitable. En réalité, cette question est assez technique et étonnamment complexe, et dépasse l'objet de ce livre[9].

Mais il faut souligner que le fait même qu'on puisse poser cette question (et y répondre) est un avantage évident de l'algorithme sur le jugement humain. D'abord, on peut évaluer les algorithmes afin de s'assurer qu'ils ne prennent pas en compte des inputs considérés comme non admissibles. On peut ensuite les tester (ou les « auditer ») pour savoir

s’ils ont un effet discriminatoire inacceptable au regard de certains critères d’équité prédéfinis. La réponse à ces questions dépendra des algorithmes et des critères retenus, mais on est sûr d’en obtenir une.

On ne peut pas en dire autant des juges humains, qu’il est beaucoup plus difficile de soumettre à ce type d’examen. Des humains peuvent être biaisés sans le savoir, et donc discriminer d’une manière dont eux-mêmes n’ont pas conscience. Un observateur extérieur, ou même le système judiciaire, peut avoir beaucoup de mal à démontrer leurs biais, car il est difficile de les tester en leur soumettant un très grand nombre de cas différents. À cet égard, les algorithmes, auxquels on reproche souvent leur opacité, peuvent être bien plus transparents que les humains[10].

En somme, il est bien sûr indispensable d’attirer l’attention sur le danger de développer des algorithmes exempts de bruit mais biaisés, tout comme sur le danger de règles exemptes de bruit mais biaisées. La vraie question, cependant, n’est pas de savoir s’il existe des algorithmes biaisés : la réponse est oui, évidemment. Elle est de savoir s’il est possible de concevoir des algorithmes supérieurs aux juges humains sur un ensemble de critères également importants : la justesse et la réduction du bruit d’une part, la non-discrimination et l’équité d’autre part.

Un grand nombre d’éléments suggèrent que, là aussi, la réponse est oui ; c’est-à-dire que les algorithmes *peuvent* être supérieurs au jugement humain sur n’importe quelle combinaison de critères retenus. (Nous soulignons *peuvent* : ce n’est pas vrai de tous les algorithmes.) Comme nous l’avons vu au chapitre 10, par exemple, pour prendre des décisions de libération conditionnelle, un algorithme peut faire des prédictions plus exactes que les juges humains, tout en produisant moins de discriminations raciales. De même, un algorithme de sélection de CV peut sélectionner un groupe de candidats non seulement meilleurs, mais aussi plus divers, que des recruteurs humains.

Ces exemples, et beaucoup d'autres, nous conduisent à une conclusion inévitable : même si, dans un monde incertain, un algorithme prédictif ne peut jamais être parfait, il peut être beaucoup moins imparfait que le jugement humain, qui est toujours sujet au bruit et souvent biaisé. Cette supériorité vaut pour la validité : un bon algorithme fait presque toujours de meilleures prédictions. Elle vaut aussi pour la non-discrimination : un bon algorithme *peut* être moins biaisé que le juge humain. Dans la mesure où cette conclusion est exacte, notre répugnance pour les algorithmes et notre préférence intuitive pour le jugement humain méritent d'être sérieusement réexaminées.

---

La question des algorithmes est particulièrement controversée, mais notre conclusion principale est plus générale – et assez simple : le bruit n'est pas une bonne chose, et cela reste vrai même si l'on estime qu'un moyen particulier de le réduire en est une mauvaise.

Il est vrai que les stratégies de réduction du bruit peuvent être coûteuses, qu'elles sont souvent difficiles à mettre en œuvre, et qu'elles peuvent produire des effets indésirables. Dans certains cas, donc, il est possible que la réduction du bruit ne se justifie pas. Mais la plupart du temps, ces inconvénients bien réels ne sont qu'un prétexte, parce qu'ils ne constituent pas une raison suffisante pour tolérer les injustices et les erreurs qu'occasionne le bruit. Et quand une mesure de réduction du bruit produit elle-même des erreurs, notamment sous forme de biais, c'est effectivement un problème sérieux. Mais dans ce cas, la solution n'est pas de renoncer à réduire le bruit : c'est de trouver de meilleures manières de le faire.

## À propos des coûts de la réduction du bruit

« Dans l'enseignement, si nous voulions éliminer totalement le bruit dans la notation, il faudrait dépenser beaucoup d'argent. En général, nous ne pouvons pas nous permettre de mobiliser plusieurs professeurs pour noter la même copie ; mais il y a sans doute des situations où l'effort se justifie. »

« Si, au lieu de se reposer sur le jugement humain, un réseau social décidait d'interdire certains mots, sans tenir compte du contexte, il éliminerait le bruit, mais créerait aussi beaucoup d'erreurs. Le remède pourrait être pire que le mal. »

« Il est vrai que certaines règles et certains algorithmes sont biaisés. Mais les humains aussi. Par ailleurs, il ne faut pas oublier que le bruit aussi cause des erreurs, et que tous les jugements humains en sont victimes. Ce que nous devrions donc nous demander, c'est comment concevoir un algorithme à la fois exempt de bruit et moins biaisé que les jugements humains. »

« Si nos efforts pour réduire le bruit ne réussissent pas du premier coup – par exemple, si nous nous retrouvons avec des directives trop rigides ou des règles qui produisent accidentellement du biais –, ce n'est pas une raison pour abandonner : il faut essayer autrement. »

CHAPITRE 27

# Une affaire de dignité

Imaginons qu’une banque vous ait refusé un prêt immobilier sans avoir analysé votre situation : votre note de crédit est inférieure au seuil minimal de la banque. Ou que vous soyez un professionnel très qualifié dans votre domaine, et que vous ayez impressionné le recruteur qui vous a fait passer un entretien, mais que votre candidature soit rejetée : il y a quinze ans de cela, vous avez été condamné pour usage de stupéfiants, et l’entreprise a pour politique de ne recruter aucun individu ayant un casier judiciaire. Imaginons, enfin, que vous ayez été accusé d’un crime et que la possibilité de rester en liberté sous caution dans l’attente de votre procès vous ait été refusée, non pas après une audience devant un juge, mais parce qu’un algorithme a calculé que les personnes qui vous ressemblent présentent un risque de non-comparution trop élevé.

Il suffit de lire ces scénarios pour les trouver problématiques. Nous considérons presque tous comme essentiel d’être traités comme des individus uniques. Nous voulons qu’un humain examine notre situation dans toute sa singularité. Nous savons, ou nous ne savons pas, que ce traitement individualisé produit du bruit, mais si c’est le prix à payer pour être traité de manière individualisée, nous pensons généralement que ce bruit est justifié. Et nous protestons chaque fois que des personnes ne sont

pas traitées, pour reprendre l'expression de la Cour suprême des États-Unis, « comme des individus uniques mais comme les membres d'une masse indifférenciée et sans visage ».

Dans le domaine judiciaire et plus généralement dans toutes les décisions qui affectent des personnes, beaucoup considèrent que seule une audience individualisée peut donner aux personnes concernées le sentiment qu'elles sont traitées comme des individus à part entière. Le droit à un procès équitable, essentiel dans un État de droit, semble incompatible avec la tyrannie des règles : il nécessite une interaction en face à face dans laquelle un être humain, habilité à faire usage de son jugement discrétionnaire, pourra prendre en considération toute une série de facteurs.

Cette défense du jugement au cas par cas repose sur de profonds fondements moraux. On la trouve dans le droit[1], la science politique, la théologie et même la littérature. *Le Marchand de Venise*, de Shakespeare, peut être lu comme un refus du silence de la règle et comme un plaidoyer pour le jugement humain. C'est le sens de ces vers célèbres où Portia fait l'éloge de la clémence :

> *La clémence ne se commande pas.*
> *Elle tombe du ciel, comme une pluie douce,*
> *sur le lieu qu'elle domine ; double bienfaisance,*
> *elle fait du bien à celui qui donne et celui qui reçoit.*
> […]
> *Elle trône dans le cœur des rois,*
> *elle est l'attribut de Dieu même ;*
> *et le pouvoir terrestre qui ressemble le plus à Dieu est*
> *celui qui tempère la justice par la clémence.*

Parce qu'elle n'est pas contrainte par des règles, la clémence est sujette au bruit. Mais le plaidoyer de Portia est pertinent dans un grand nombre de

situations et d'organisations. Quand un employé réclame une prime, quand un particulier sollicite un prêt, quand un étudiant postule à l'université, ceux qui décident rejettent souvent certaines stratégies de réduction du bruit, et notamment les règles trop rigides. Peut-être est-ce parce qu'ils pensent, avec Portia, que la clémence ne se commande pas. Même s'ils sont conscients que leur approche est sujette au bruit [2], ils préféreront souvent s'y tenir, afin de donner aux employés, aux emprunteurs et aux étudiants le sentiment d'être écoutés et traités avec respect.

Certaines stratégies de réduction du bruit ne prêtent pas le flanc à cette critique. Quand la décision est prise par trois personnes au lieu d'une, l'écoute individualisée reste possible. Quand des directives sont mises en place, elles peuvent encadrer la décision tout en laissant aux décideurs un pouvoir discrétionnaire important. Mais certaines mesures de réduction du bruit, et notamment les règles rigides, suppriment ce pouvoir, et peuvent amener certains à conclure que le processus attente à leur dignité.

Faut-il leur donner raison ? Il est vrai que nous voulons tous pouvoir bénéficier d'une écoute individualisée. La possibilité d'être entendu revêt une valeur humaine incontestable. Mais si l'écoute individualisée produit des erreurs qui ont des conséquences tragiques, des injustices, des inégalités, et des coûts bien plus élevés, alors son éloge devrait être pour le moins nuancé. Ici encore, la meilleure réponse n'est pas de renoncer à réduire le bruit, mais de chercher une meilleure manière de le faire. Et si la meilleure stratégie de réduction du bruit possible produit moins d'erreurs, il est possible que les personnes qui en bénéficient – justiciables, patients, étudiants ou employés – la trouvent préférable au traitement individualisé, même si elle réduit ou supprime la possibilité d'être entendu. Encore faut-il pour cela qu'elles soient conscientes que le traitement au cas par cas produit beaucoup de bruit, c'est-à-dire des erreurs graves et des injustices flagrantes.

## *Des règles inflexibles face à des valeurs changeantes ?*

Imaginons qu'une institution réussisse à éliminer le bruit. Qu'une université, par exemple, définisse la *faute grave* avec suffisamment de précision pour que tout membre de la faculté et tout étudiant sachent en quoi elle consiste et ne consiste pas. Ou qu'une grande entreprise définisse précisément ce que signifie la *corruption*, afin que nul employé de l'entreprise n'ignore ce qui est permis et ce qui est interdit. Ou encore que cette entreprise réduise le bruit de façon significative dans ses décisions de recrutement, par exemple en décidant qu'elle ne recrutera que des candidats qui ont obtenu certains diplômes.

Que se passe-t-il, alors, quand les valeurs et les priorités de ces organisations changent ? Certaines stratégies de réduction du bruit semblent incapables de s'adapter à ces évolutions, et cette inflexibilité peut poser un sérieux problème, étroitement lié à celui du traitement individualisé et de la dignité humaine.

Une décision aussi célèbre qu'énigmatique de la Cour suprême des États-Unis, rendue en 1974, illustre ce problème[3]. L'affaire concernait un réseau d'écoles qui exigeait de toutes ses enseignantes enceintes qu'elles prennent un congé non rémunéré de cinq mois avant la date attendue de la naissance de l'enfant. Jo Carol LaFleur, une enseignante, estimait qu'elle était parfaitement capable de faire son travail, que la règle était discriminatoire et que la durée de cinq mois était excessive.

La Cour suprême donna raison à l'enseignante. Elle n'invoqua cependant pas la discrimination, ni le caractère excessif de la durée de cinq mois. L'argument sur lequel elle fonda son jugement fut tout autre : la Cour jugea anticonstitutionnel le fait que l'employeur de Jo Carol LaFleur ne lui ait pas donné la possibilité de démontrer que, pour elle en particulier, un congé était inutile. Dans les termes de son arrêt,

*Il n'y a pas eu d'appréciation individualisée par le médecin de l'enseignante, ni par celui de l'employeur, de la capacité de l'enseignante à continuer son travail. Les règles contiennent une présomption irréfragable d'incapacité physique, et cette présomption s'appliquerait même en présence d'expertises médicales réalisées sur une personne particulière et qui démontreraient le contraire.*

Un congé obligatoire de cinq mois pour une enseignante semble manifestement absurde. La Cour ne retint pourtant pas ce point. Elle choisit plutôt de dénoncer le caractère irréfragable de la présomption et l'absence d'« appréciation individualisée ». Il semble donc que pour la Cour, comme pour Portia, la clémence ne se commande pas : il fallait impérativement que quelqu'un puisse prendre en considération la situation particulière de Jo LaFleur.

Le problème de ce raisonnement est que, sans hygiène de la décision, le bruit est garanti. Qui doit décider si Jo LaFleur est apte à travailler ? La décision sera-t-elle la même pour d'autres femmes se trouvant dans la même situation ? En tout état de cause, un grand nombre de règles reposent sur une présomption irréfragable. Le fait de dépasser la limitation de vitesse constitue-t-il une présomption irréfragable de conduite dangereuse ? *Quid* d'un âge minimum pour boire de l'alcool ou pour voter ? De l'interdiction de conduire en état d'ivresse ? En invoquant des exemples de ce genre, d'aucuns ont objecté qu'il serait intenable de bannir par principe les « présomptions irréfragables », qui ont pour objet et pour effet de réduire le bruit.

Des commentateurs influents à l'époque ont pourtant défendu la décision de la Cour. Leur argument central était que les valeurs de la société changent avec le temps, et qu'il faut donc éviter d'édicter des règles rigides qui figent ce qui est permis ou interdit[4]. Notant en particulier que les normes sociales quant au rôle des femmes dans la société évoluaient vite, ils

estimaient qu'une appréciation individualisée de chaque situation était mieux à même d'intégrer ces normes changeantes. Un système de règles strictes aurait peut-être l'avantage d'éliminer le bruit, mais il aurait aussi l'inconvénient de figer les normes et les valeurs existantes.

Au-delà du cas de Jo LaFleur, on peut estimer qu'un système sujet au bruit est plus flexible, en ce sens qu'il s'adapte plus facilement aux changements de son environnement. À mesure que les valeurs de la société évoluent, par exemple, les juges peuvent faire usage de leur pouvoir d'appréciation discrétionnaire pour prononcer des peines moins lourdes dans les affaires de consommation de stupéfiants, ou des peines plus élevées dans les affaires de viol. Indéniablement, si certains juges suivent ces évolutions plus vite que d'autres, cette flexibilité se traduira par un certain degré d'injustice, puisque des individus placés dans la même situation au même moment seront traités différemment. Mais cette injustice peut permettre au système dans son ensemble de s'adapter à un contexte sociétal changeant.

La même logique peut être invoquée bien au-delà du système judiciaire. Les entreprises peuvent par exemple choisir de laisser un certain flou dans leurs politiques afin de les faire plus facilement évoluer. En voici un exemple personnel : quand l'un des auteurs de ce livre a rejoint, il y a de nombreuses années, un grand cabinet de conseil, le guide de bienvenue (déjà ancien) qui lui fut remis spécifiait avec moult détails les frais professionnels considérés comme acceptables. Par exemple : « en voyage, un seul coup de téléphone chez vous pour dire que vous êtes bien arrivé ; des frais de pressing pour un costume ; des pourboires pour les grooms ». Les règles étaient exemptes de bruit mais, même pour l'époque, manifestement désuètes (et sexistes). Elles ne tardèrent pas à être remplacées par des principes plus généraux, et donc plus aptes à évoluer dans le temps. Ainsi, les dépenses doivent désormais être « adaptées et raisonnables ».

Cette défense du bruit au nom de la flexibilité peut être sensée, mais elle appelle plusieurs réponses. D'abord, répétons-le : la critique d'une stratégie de réduction du bruit ne justifie pas le bruit. Il existe beaucoup de stratégies de réduction du bruit qui ne soulèvent pas l'objection d'inflexibilité, parce qu'elles n'entravent en rien la capacité du système à évoluer – au contraire. L'agrégation de jugements multiples, par exemple, permet aux jugements d'évoluer à mesure que les juges adoptent de nouvelles valeurs. De même, une échelle basée sur une vision externe peut s'adapter aux mutations de son environnement, puisqu'elle utilise comme cadre de référence un ensemble de cas récents ou contemporains.

En tout état de cause, même si l'on utilise des règles ou des directives pour réduire le bruit, rien ne dit que celles-ci doivent être éternelles. Elles peuvent être révisées systématiquement chaque année. On peut parfaitement décider qu'en raison d'événements importants ou de valeurs et croyances nouvelles, les règles doivent être modifiées. Dans le système pénal, par exemple, le législateur peut réduire les peines pour certains crimes ou délits et les durcir pour d'autres. Il peut décriminaliser certains actes et en criminaliser d'autres qui étaient considérés auparavant comme acceptables. Dans les relations entre clients et fournisseurs, des pratiques qui étaient naguère considérées comme normales ont cessé de l'être, et dans beaucoup d'entreprises, les codes de conduite ont évolué pour refléter ces changements.

En résumé, un peu de bruit dans un système peut permettre que des valeurs sociétales en constante évolution soient intégrées dans les jugements que produit ce système ; et, toutes choses égales par ailleurs, cela peut être une bonne chose. Mais on peut réduire considérablement le bruit sans empêcher la bonne prise en compte de ces évolutions. En outre, dans de nombreux domaines, cet argument n'a aucune pertinence. Si deux clients qui se plaignent d'un ordinateur défectueux sont traités différemment par le fabricant, ou si deux individus souffrant de la même pathologie font l'objet

de diagnostics médicaux différents, cette incohérence n'a rien à voir avec l'évolution des valeurs de la société.

## *Des règles trop précises, donc trop faciles à contourner ?*

Le propre d'un système sujet au bruit est de laisser à ceux qui formulent des jugements la possibilité de s'adapter à toutes les situations, même les plus inattendues. Si on les en empêche, il peut devenir plus facile de prévoir leurs décisions et donc de trouver des manières de les contourner. Un des arguments avancés pour tolérer le bruit est donc qu'il peut être nécessaire pour permettre aux institutions publiques et privées d'éviter ce genre de contournement.

Le Code des impôts en fournit un bon exemple. Dans un monde idéal, les impôts et taxes dont vous êtes redevable ne devraient pas être sujets au bruit. La loi devrait être claire et prévisible, et des contribuables dont les situations sont identiques ne devraient pas être traités différemment. Mais si les règles qui gouvernent la fiscalité sont si précises qu'elles éliminent le bruit, certains contribuables trouveront inévitablement un moyen de s'y soustraire. La question fait débat chez les fiscalistes de tous pays : vaut-il mieux édicter des règles très détaillées qui éliminent le bruit, ou au contraire conserver une certaine imprécision, qui rend l'impôt moins prévisible, mais qui limite les possibilités de montages opportunistes, conformes à la lettre de la loi mais contraires à son esprit ?

De la même manière, beaucoup d'entreprises, d'universités ou d'autres organisations ont des codes de conduite qui interdisent les « comportements inappropriés » sans plus de précision. Il s'ensuit inévitablement du bruit, ce qui ne peut pas être une bonne chose. Mais si l'alternative est de dresser une

liste exhaustive de tout ce qui est interdit, alors tout comportement qui n'y figure pas, si néfaste soit-il, sera présumé toléré – ce qui n'est pas bon non plus.

Les règles, par définition, ont des limites précises. Il est donc souvent possible de les contourner en agissant d'une manière qui n'est pas techniquement interdite, mais qui peut s'avérer tout aussi problématique. (Tous les parents d'adolescents en ont fait l'expérience.) D'où une objection aux règles en tous genres : puisqu'il est impossible d'élaborer des règles prohibant tout ce qui devrait l'être, il ne reste plus qu'à tolérer le bruit.

Il est vrai que dans certaines situations, comme par exemple en matière de fiscalité, les règles clairement définies font naître un risque de contournement. Ce risque pourrait être une bonne raison d'adopter d'autres stratégies de réduction du bruit, comme l'agrégation des jugements, voire peut-être de tolérer un peu de bruit. Mais le mot « pourrait » est crucial. La simple possibilité que des règles soient contournées ne suffit pas à discréditer l'idée même de règle. L'objection n'est donc pas une objection de principe, mais une difficulté de mise en œuvre, à traiter avec pragmatisme. En pratique, il faut évaluer l'ampleur des abus que permet le contournement des règles, mais aussi le niveau de bruit qu'on observe (ou qu'on craint) en l'absence de ces mêmes règles. S'il n'y a que peu d'abus, mais beaucoup de bruit, alors il peut être préférable de réduire le bruit. Nous reviendrons sur cette question au chapitre 28.

## *Les vertus dissuasives de l'imprévisibilité ?*

Imaginons que votre objectif soit de dissuader des employés, des étudiants, ou des citoyens ordinaires de mal se comporter, dans un domaine ou dans

un autre. Un peu d'imprévisibilité ne nuit pas forcément. L'employeur pourrait se dire : « Si les employés savent que ce comportement constitue une faute, mais ne savent pas s'ils risquent un avertissement, une mise à pied, ou un licenciement immédiat, alors ils s'en abstiendront. » Et en matière de justice pénale, le législateur pourrait se dire : « Ce n'est pas un problème que les futurs délinquants ne soient pas sûrs de la sanction qu'ils encourent. Si le fait que leur peine dépende d'une loterie peut les dissuader de franchir la ligne rouge, le bruit qui en résulte n'est pas un problème, au contraire. »

En théorie, ces arguments ne peuvent pas être écartés, mais ils ne sont pas très convaincants. Ce qui semble importer, c'est la valeur attendue de la sanction, et avoir une chance sur deux d'écoper d'une amende de 5 000 dollars équivaut, de ce point de vue, à la certitude d'une amende de 2 500 dollars. Bien sûr, le scénario le plus défavorable sera plus visible pour ceux qui ont une forte aversion au risque. Ceux-là pourraient être plus fortement dissuadés par les 50 % de chances d'avoir une amende de 5 000 dollars que par 100 % de chances d'en devoir 2 500. Mais ce ne sera pas le cas des individus qui ont, au contraire, un certain appétit pour le risque.

Pour savoir si un système sujet au bruit, donc imprévisible, est plus ou moins dissuasif, il faut donc d'abord savoir si les contrevenants potentiels qu'on cherche à dissuader fuient le risque ou le recherchent. Et, en tout état de cause, si l'objectif est d'être dissuasif, ne vaudrait-il pas mieux durcir la peine, tout en augmentant la probabilité qu'elle soit effective ? Cette réduction du bruit accroîtrait la dissuasion tout en éliminant de l'injustice.

## *Le bruit, condition de la créativité et de la motivation ?*

Certains efforts de réduction du bruit nuiraient-ils à la motivation et à l'engagement ? Pourraient-ils affecter la créativité et empêcher des avancées décisives ? C'est ce que semblent penser certaines organisations. Parfois, sans doute, elles ont raison ; mais ici encore, il faut d'abord préciser à quelle stratégie de réduction du bruit elles s'opposent.

Rappelez-vous la réaction extrêmement négative de nombreux juges contre les directives en matière de peine[5]. Comme l'écrivait l'un d'entre eux, « il nous faut de nouveau apprendre à faire confiance à l'exercice du jugement dans un tribunal ». De façon générale, les personnes en position d'autorité tiennent à leur pouvoir d'appréciation discrétionnaire. La possibilité d'exercer leur jugement a pour elles une grande valeur et fait partie de leur identité professionnelle. Si on leur retire cette possibilité, elles peuvent avoir l'impression qu'on ne leur fait pas confiance, qu'on leur lie les mains, voire qu'on les humilie. Par conséquent, le risque est réel que certains de ces professionnels hautement qualifiés se rebellent.

Pour le dire plus positivement, un système sujet au bruit peut avoir l'avantage de préserver la motivation des professionnels qui décident comme ils l'entendent. Si des agents du service des réclamations peuvent répondre comme ils le jugent bon aux critiques des clients, si des managers peuvent évaluer leurs subordonnés à leur guise, et si les souscripteurs d'assurances peuvent fixer les niveaux de prime qui leur semblent optimaux, sans doute tous ces employés apprécieront-ils davantage leur travail. Inversement, si l'on prend des mesures pour réduire le bruit dans leurs jugements, ils pourront avoir l'impression que leur capacité d'action a été diminuée, et que leur travail est devenu mécanique, voire robotique.

Les organisations peuvent choisir de tenir compte de ces sentiments, d'abord parce qu'elles y sont sensibles, mais aussi parce que c'est dans leur intérêt. Sans motivation, sans implication, et sans possibilité d'exercer son jugement et de développer ses compétences, il est difficile de trouver des

idées nouvelles. En ce sens, l'excès de règles pourrait nuire à l'inventivité et à l'ingéniosité, et donc mettre en péril l'avenir de l'organisation.

Ces objections à la réduction du bruit s'appliquent sans doute à un grand nombre de professionnels, mais pas à tous. Des tâches de jugement différentes doivent, à cet égard, être évaluées différemment. Le diagnostic de l'angine ou celui de l'hypertension, par exemple, ne sont sans doute pas de bons terrains pour l'exercice d'une capacité de jugement discrétionnaire : si le jugement diagnostique est sujet au bruit, cela ne doit rien à la créativité des médecins.

Mais surtout, il devrait être possible de réduire le bruit tout en préservant la motivation des individus et en restant réceptif à leurs suggestions. C'est précisément le but de certaines stratégies de réduction du bruit – la structuration des jugements complexes, par exemple. Une organisation soucieuse de réduire le bruit tout en conservant un bon état d'esprit devra donc choisir des techniques d'hygiène de la décision qui lui permettent de concilier ces objectifs. Et si des règles sont mises en place, et limitent donc l'exercice du jugement individuel, les professionnels de l'organisation peuvent et doivent participer au processus qui permet de définir ces règles, de les contester éventuellement, de les faire évoluer, et d'y faire des exceptions raisonnées – sans pour autant s'y soustraire au cas par cas en exerçant un jugement discrétionnaire.

Dans plusieurs ouvrages roboratifs, le juriste Philip Howard avance des arguments similaires en faveur d'une plus grande souplesse de jugement[6]. Il plaide pour que les lois prennent la forme non pas de règles prescriptives, qui éliminent le bruit, mais de principes généraux : « être raisonnable », « agir prudemment », « ne pas prendre de risques excessifs ».

Pour Howard, l'univers de la réglementation publique est devenu fou parce qu'il est trop rigide. Enseignants, agriculteurs, promoteurs, médecins : tous ces professionnels, et beaucoup d'autres, croulent sous des règles qui leur disent ce qu'ils doivent faire et comment le faire. Howard pense qu'il

vaudrait mieux permettre aux gens d’user de leur créativité pour décider eux-mêmes de la meilleure manière d’atteindre des objectifs pertinents, qu’il s’agisse de meilleurs résultats scolaires, d’une eau plus propre, d’un nombre moindre d’accidents de chantiers, ou de patients en meilleure santé.

Certains arguments d’Howard font mouche, mais il est important de s’interroger sur les conséquences qu’aurait la généralisation des principes qu’il propose, notamment en termes d’aggravation du bruit et du biais. Personne, certes, ne se présente, dans l’absolu, comme un défenseur de la rigidité, mais il arrive qu’elle soit le meilleur moyen de réduire le bruit et le biais, et donc de diminuer l’erreur. Plus les principes sont généraux, et plus il faut s’attendre à ce qu’il y ait du bruit dans leur interprétation et dans leur mise en œuvre, avec des conséquences qui peuvent être intolérables.

---

Les craintes que suscite l’idée de réduire le bruit sont parfois fondées, car il est indéniable que certaines mesures de réduction du bruit ont des coûts non nuls et des effets indésirables bien réels. La dignité des personnes, la capacité des normes à évoluer, la dissuasion des contrevenants, la flexibilité face aux abus, la motivation des professionnels : toutes ces considérations sont importantes.

Mais des arguments contre telle ou telle technique de réduction du bruit, pour convaincants qu’ils soient, ne font pas une défense du bruit lui-même. Si certaines techniques d’hygiène de la décision ne sont pas adaptées à une situation ou à une autre, notre réaction ne devrait pas être de nous résigner au bruit. Elle devrait être de chercher des stratégies pour le réduire, sans pour autant compromettre d’autres valeurs importantes.

## À propos de la dignité

« Ne réduisez pas le bruit ! Les gens veulent qu'un être humain les entende et ait le pouvoir d'améliorer les choses. C'est une affaire de dignité, et la dignité n'a pas de prix. »

« Ne réduisez pas le bruit ! Les valeurs morales ne cessent d'évoluer. Si vous cadenassez tout, vous empêcherez toute évolution. »

« Ne réduisez pas le bruit ! Un peu d'incertitude, donc de bruit, renforce la dissuasion. »

« Ne réduisez pas le bruit ! Vous vous retrouverez avec des règles claires, mais certains trouveront le moyen de les contourner. »

« Ne réduisez pas le bruit ! Les gens ne sont pas des robots. Trop encadrés, ils se démotiveront et n'apporteront plus d'idées originales. »

« En vérité, tous ces arguments ne sont pas convaincants. Ou plutôt, ils sont convaincants en tant que critiques de certaines stratégies de réduction du bruit, mais ils ne justifient pas le bruit lui-même. On peut respecter la dignité des personnes, laisser de la place pour que les valeurs morales puissent évoluer et permettre aux gens d'exprimer leur créativité, sans tolérer pour autant l'injustice et le coût du bruit. »

CHAPITRE 28

# Des règles ou des principes ?

Pour réduire le bruit, ou pour décider si et comment le faire, on distinguera utilement deux modes de régulation des comportements : la règle, précise, et le principe, plus général. Les organisations choisissent souvent l'une ou l'autre, ou une combinaison des deux.

Une entreprise peut par exemple décider que ses salariés travailleront à certaines heures, que leurs congés ne peuvent pas dépasser deux semaines consécutives, et que quiconque fait fuiter des informations dans la presse sera licencié. Ce faisant, elle édicte des règles. Mais elle peut aussi (du moins, dans certains pays) fixer des principes plus généraux, par exemple que les employés feront « une journée de travail raisonnable », que les vacances seront décidées « au cas par cas en fonction des besoins de l'entreprise » et que les fuites dans la presse « seront sanctionnées comme il se doit ».

En droit, des règles peuvent stipuler qu'il est interdit de rouler au-dessus d'une certaine vitesse, que des ouvriers ne doivent pas être exposés à des substances cancérogènes, ou que la notice d'un médicament doit comprendre des informations détaillées sur la nature et la fréquence de ses effets secondaires et sur ses contre-indications. Des principes, en revanche, se contenteraient d'énoncer qu'on est tenu de conduire « prudemment »,

que les employeurs doivent préserver la sécurité et la santé des employés « dans toute la mesure du possible », ou que les sociétés pharmaceutiques doivent décider « raisonnablement » de ce qui constitue une information « suffisante ».

Ces exemples illustrent la distinction centrale entre règles et principes. La règle est conçue pour éliminer toute possibilité d'appréciation par les personnes qui devront l'appliquer ; le principe, au contraire, pour leur en laisser une. Dès lors qu'on définit des règles, le bruit sera sérieusement réduit, car les personnes qui appliquent les règles n'ont plus qu'à répondre à une question factuelle : à quelle vitesse allait la voiture ? Ce travailleur a-t-il été exposé à des substances cancérogènes ? Ce médicament donne-t-il toutes les informations requises sur les contre-indications ?

Cela ne signifie pas que les règles fassent totalement disparaître le bruit, car la recherche des faits peut elle-même nécessiter un jugement, sujet au bruit comme au biais. Nous en avons vu de nombreux exemples. Mais les auteurs de règles cherchent à réduire ce risque. À l'extrême, quand une règle se résume à un chiffre (« la vitesse limite est de 80 km/h »), le bruit est très fortement réduit. La caractéristique centrale des règles est donc qu'elles *limitent le rôle du jugement*. Les juges (au sens large que nous donnons ici à ce terme) ont moins à faire. Ils appliquent les règles. Qu'on le déplore ou qu'on s'en réjouisse, ils ont moins de marge de manœuvre.

Les principes sont tout à fait différents. Là où prévalent des principes, les juges ont fort à faire pour en préciser le sens. Il faut des jugements pour déterminer, par exemple, ce que signifient les mots « raisonnablement » ou « dans la mesure du possible ». Non contents d'établir des faits, les juges doivent donner un contenu à des formules relativement vagues, parce que les auteurs des principes leur ont, de fait, sous-traité cette responsabilité. Celui qui fixe une règle conserve le pouvoir, celui qui définit un principe le délègue en partie.

Les directives évoquées au chapitre 22 peuvent être plus proches des règles ou des principes. Le score d'Apgar, par exemple, n'est pas une règle, puisqu'il demande encore du jugement, mais il encadre néanmoins le diagnostic assez précisément. Quand les directives sont si précises qu'elles empêchent tout exercice du jugement, elles deviennent des règles : ainsi, les algorithmes fonctionnent comme des règles, pas comme des principes.

## *Divisions et ignorance*

Il faut rappeler, pour commencer, que quand une organisation, un groupe ou une société sont divisés, il y sera beaucoup plus facile de se mettre d'accord sur des principes que sur des règles. Des dirigeants d'entreprise peuvent être d'accord pour bannir les comportements inappropriés sans savoir précisément s'ils sont d'accord sur le sens qu'ils donnent à ces termes. Tous ceux qui condamnent le harcèlement sexuel sur le lieu de travail ne sont pas forcément d'accord sur les actes et les propos qu'il recouvre. Une université peut interdire le plagiat sans préciser exactement où il commence. Des citoyens peuvent s'accorder sur l'importance vitale de protéger la liberté d'expression sans préciser si celle-ci s'étend à la publicité commerciale, aux propos haineux ou aux obscénités. Et ils peuvent souhaiter que le législateur adopte des règles « prudentes » pour réduire les émissions de gaz à effet de serre sans avoir la moindre idée de ce que signifie « prudentes ». Les constitutions énoncent un grand nombre de principes (par exemple, la liberté de conscience), comme le fait aussi la Déclaration universelle des droits de l'homme (« Tous les êtres humains naissent libres et égaux en dignité et en droits »).

Ces exemples le montrent : adopter des principes sans les définir de façon précise a de fortes chances de produire du bruit, mais on devra pourtant souvent s'en contenter, faute d'accord sur des règles. Quand des

législateurs parviennent à un compromis sur des principes, le bruit qui en résulte est en quelque sorte le prix à payer pour qu'une loi soit adoptée. De même, les médecins peuvent se mettre d'accord sur des principes de bonnes pratiques, ou les assureurs sur les principes qui gouvernent la fixation des primes d'assurance, alors que des tentatives pour élaborer des règles précises se heurteraient à des désaccords insurmontables.

Mais les divisions sociales et politiques ne sont pas la seule raison de préférer les principes. Quelquefois, le vrai problème est que nous ne disposons pas des informations nécessaires pour édicter des règles adéquates. Une université peut être incapable de définir précisément les règles qui doivent déterminer quels professeurs seront promus. Un employeur peut avoir du mal à prévoir toutes les circonstances justifiant de sanctionner ses employés. Des parlementaires peuvent ne pas connaître le seuil tolérable de pollution de l'air par diverses particules, l'ozone, le dioxyde d'azote ou le plomb. Dans tous ces cas, la meilleure stratégie est de fixer des principes et de se reposer sur des experts de confiance pour en préciser la signification, même si cette démarche a pour effet de produire du bruit.

Il faut souligner que les règles, si elles éliminent le bruit, peuvent être biaisées de bien des façons. Une règle peut interdire aux femmes de devenir agents de police, ou exclure les Irlandais. Si ces règles sont respectées, il n'y aura pas de bruit, mais il y aura évidemment des biais.

## *Les chefs et le contrôle des subordonnés*

Le choix entre règles et principes est d'une importance capitale pour toutes les organisations, publiques ou privées. Il se présente chaque fois qu'un « principal » veut contrôler un « agent » auquel il a délégué des tâches. Par exemple, comme on l'a vu au chapitre 2, les souscripteurs d'assurances

s’efforcent de trouver le niveau de prime idéal (ni trop élevé, ni trop bas), pour que leur entreprise maximise ses profits. Pour les guider, leurs supérieurs doivent-ils leur donner des règles ou se contenter de fixer des principes ? Plus généralement, quand un manager doit-il donner à ses subordonnés des instructions précises, plutôt que des principes généraux (« fais usage de ton meilleur jugement ») ? Et quand un médecin doit-il prescrire à son patient de « prendre un cachet matin, midi et soir » (une règle), plutôt que « en cas de besoin » (un principe) ?

Nous avons vu qu’un réseau social comme Facebook est obligé de se préoccuper du bruit et des moyens de le réduire. L’entreprise peut enjoindre à ses employés de retirer le contenu de tout message contrevenant à une règle simple (interdire la nudité, par exemple), ou leur demander d’appliquer des principes (interdire tout contenu violent). Les *community standards* de Facebook (« normes de la communauté »), rendus publics pour la première fois en 2018, sont un mélange fascinant de règles et de principes, par ailleurs très nombreux. Après leur publication, beaucoup d’utilisateurs de Facebook se sont soulevés contre ce texte. Comme des milliers de modérateurs allaient devoir produire des jugements pour interpréter ces principes, expliquaient-ils, leurs décisions de retirer ou non les messages examinés seraient forcément très variables, et produiraient inévitablement un bruit excessif : des injustices et des erreurs. Pour comprendre pourquoi cette variabilité était inévitable, citons un extrait des *community standards* de Facebook de 2020[1] :

> *Nous définissons la parole haineuse comme une attaque directe contre des personnes, fondée sur ce que nous appelons des caractéristiques protégées : la race, l’appartenance ethnique, l’origine nationale, la religion, l’orientation sexuelle, la caste, le sexe, le genre, l’identité, une maladie ou un handicap grave. Nous définissons l’attaque comme un*

*discours violent ou déshumanisant, une déclaration d'infériorité, ou un appel à l'exclusion ou à la ségrégation.*

Dans l'application d'une définition comme celle-ci, les modérateurs seront inévitablement sujets au bruit. Qu'est-ce qu'« un discours violent ou déshumanisant » ? N'ignorant pas ces questions, Facebook décida d'y répondre en allant dans le sens de règles tranchées, précisément en vue de réduire le bruit. Ces règles furent rangées dans un document non publié, les *implementation standards* [« normes d'application »], d'une douzaine de milliers de mots, que le magazine *The New Yorker* s'est procuré[2]. Dans le document public, les « normes de la communauté », l'article régissant le contenu mis en ligne, image ou texte, commençait par un principe : « Nous retirons tout contenu glorifiant la violence. » (Qu'est-ce que cela signifie exactement ?) Les « normes d'application », elles, dressent une liste précise des images que les modérateurs de contenu doivent retirer. Sont par exemple citées les images « de corps humains brûlés ou calcinés » ou le « détachement de parties non régénérables du corps humain ». Pour résumer et simplifier, les *community standards* ressemblent plutôt à des principes, les *implementation standards* plutôt à des règles.

Dans le même esprit, une compagnie aérienne peut demander à ses pilotes d'obéir à des règles ou de respecter des principes – pour savoir, par exemple, s'il faut revenir à l'aérogare au bout de quatre-vingt-dix minutes passées sur le tarmac, ou quand exactement allumer le voyant « attachez vos ceintures ». La compagnie peut être tentée d'édicter des règles qui limitent les marges d'appréciation des pilotes. Mais elle peut aussi penser que les pilotes, dans certaines circonstances, doivent exercer leur jugement, et que les principes sont en ce sens préférables.

Dans ce genre de situation, le bruit est souvent le résultat de principes trop vagues. Quand ceux-ci ont laissé le champ libre au bruit, et quand tout le monde peut observer que des individus placés dans la même situation ne

sont pas traités de la même façon, on observe souvent un mouvement pour demander des règles. Comme dans le cas de la réforme de la peine pénale, ce mouvement peut prendre de l'ampleur, et se déclenche généralement à la suite d'une forme d'audit de bruit.

## *Le retour du refoulé*

Réfléchissons un instant à cette question : quelles sont les personnes qui peuvent être considérées comme des personnes handicapées, au sens où cela leur donnerait droit aux aides réservées aux personnes en situation d'incapacité de travail ? Si la question est formulée en ces termes (un principe), des juges prendront des décisions *ad hoc* qui seront sujettes au bruit et par conséquent injustes. C'était naguère le cas aux États-Unis, et les disparités de traitement qui en résultaient étaient choquantes. Des individus souffrant de la même manière – de dépression grave, par exemple – y étaient traités de façon différente. Pour y remédier, les responsables publics ont adopté quelque chose qui ressemble davantage à une règle : une *matrice de handicap*. Cette matrice repose sur des jugements relativement mécaniques prenant en compte le niveau d'éducation de la personne, sa situation géographique et ses capacités physiques restées intactes. L'objectif est de réduire le bruit dans les décisions.

La principale réflexion sur ce problème est celle d'un professeur de droit, Jerry Mashaw, qui qualifie de *justice bureaucratique* ces efforts pour éliminer le bruit[3]. La formule mérite d'être retenue. Mashaw considère la création de la matrice comme fondamentalement juste, précisément parce qu'elle promet d'éliminer le bruit.

Dans certaines situations, cependant, la promesse de justice bureaucratique pourrait ne pas se réaliser. Même quand une institution

décide d’adopter des décisions contraintes par des règles, le risque est grand que du bruit réapparaisse.

Imaginons que les règles produisent de très mauvais résultats dans certains cas particuliers. Si tel est le cas, les personnes chargées de les appliquer pourraient préférer les ignorer ou les contourner. De fait, ils pourraient reconquérir leur pouvoir d’appréciation discrétionnaire en recourant à une sorte de désobéissance civile discrète, difficile à contrôler ou même à repérer. Dans bien des entreprises privées, les employés ignorent les règles contraignantes qui leur semblent stupides. De même, les agences publiques peuvent refuser d’appliquer des règles qui leur semblent trop rigides. En droit pénal américain, il arrive que des procès se concluent par une *jury nullification* (invalidation par le jury) : le jury considère le prévenu comme coupable mais le déclare néanmoins innocent, refusant ainsi de se conformer à une loi qui lui semble déraisonnablement dure ou rigide.

Comme ces exemples le suggèrent, une institution qui essaie de contrôler le bruit en fixant des règles précises doit toujours veiller à ce que ces règles n’aient pas pour conséquence un retour clandestin de l’appréciation discrétionnaire. Un autre exemple frappant de ce phénomène est la réaction suscitée par la « loi des trois prises » : confrontés à la troisième infraction d’un récidiviste, de nombreux procureurs retenaient une qualification pénale moindre, afin que le délit ne rentre pas dans le champ d’application de la loi entraînant une condamnation à vie automatique. Il est évidemment paradoxal qu’une loi visant à durcir les peines pousse au contraire les procureurs à déqualifier des infractions graves. Pourtant, de tels contournements sont à la fois entièrement prévisibles et extrêmement difficiles à éviter, et même à détecter.

Quand le bruit est ainsi refoulé, personne ne peut l’entendre. Les organisations qui mettent en place des règles doivent donc surveiller leur application pour s’assurer qu’elles fonctionnent comme prévu. Dans le cas contraire, les règles devront être révisées.

## *Un cadre de réflexion*

Dans les entreprises comme dans l'administration, le choix entre règles et principes se fait souvent intuitivement, et généralement sans cadre de réflexion prédéfini ; mais il est possible de le faire de façon plus ordonnée. En première approximation, il ne dépend que de deux choses : le coût de la décision et le coût de l'erreur.

Avec de simples principes, le coût de la décision peut être élevé, parce que ceux qui décident doivent, au cas par cas, donner un contenu aux principes. Or tout jugement demande un effort. Si l'on demande aux policiers de déterminer qui roule « dangereusement vite », ils auront plus de travail que si la règle fixe la vitesse limite à 80 km/h. Avec des règles, le coût de la décision est généralement beaucoup moins élevé.

Mais le problème du coût de la décision est plus compliqué que cela. Une règle est simple à appliquer une fois en place, mais au préalable, quelqu'un doit produire une règle pertinente, et cela peut être difficile et coûteux. C'est pourquoi les systèmes judiciaires et les entreprises privées, comme d'ailleurs les normes en vigueur dans la médecine ou dans l'ingénierie, utilisent si souvent des mots comme « raisonnable », « prudent » ou « faisable ».

L'autre terme de l'équation renvoie au coût de l'erreur, c'est-à-dire au nombre d'erreurs et à leurs conséquences. Une question centrale ici est celle de la confiance que le « principal » accorde à son « agent ». Quand les agents sont compétents, fiables et bien informés, et quand ils pratiquent eux-mêmes une forme d'hygiène de la décision, des principes pourront fonctionner tout aussi bien que des règles – et ne pas produire beaucoup plus de bruit. Au contraire, le principal doit imposer des règles quand il a de bonnes raisons de ne pas faire confiance à son agent. Si l'agent est incompétent ou biaisé, et s'il ne lui est pas possible de pratiquer quelque forme d'hygiène de la décision, alors il est sain d'encadrer par des règles

son pouvoir d’appréciation discrétionnaire. Les organisations avisées savent faire varier le degré d’autonomie de leurs agents en fonction de la confiance qu’elles leur accordent.

Bien sûr, toutes les erreurs ne se valent pas. Si les agents chargés de mettre en œuvre un principe ne sont pas fiables, il y aura de nombreuses erreurs, mais si ces erreurs sont sans conséquence sérieuses, elles pourront être tolérées. Inversement, une règle peut ne produire que très peu d’erreurs, mais si celles-ci sont catastrophiques, on pourra lui préférer un principe. En somme, il n’y a pas de raison *générale* de penser que le coût de l’erreur est plus élevé avec une règle qu’avec un principe. Si une règle est parfaite, bien sûr, elle ne produira pas d’erreurs, mais les règles sont rarement parfaites.

Imaginons une loi qui interdit d’acheter de l’alcool avant 21 ans. La loi vise à protéger les jeunes gens des différents risques associés à la consommation d’alcool. En ce sens, elle va produire un grand nombre d’erreurs. Certains individus de 20, 19, 18 ou même 17 ans peuvent tout aussi bien supporter l’alcool que des individus plus âgés. D’autres, âgés de 22, 42 ou 62 ans, en sont incapables. Un principe produirait moins d’erreurs si nous pouvions trouver les mots qui conviennent et si les gens pouvaient l’appliquer correctement. Bien sûr, cela est très difficile à faire, et c’est pourquoi la consommation d’alcool est presque toujours régie par des règles simples reposant sur un critère d’âge.

Plus généralement, comme l’illustre cet exemple, la répétition en très grand nombre des décisions plaide pour la définition de règles claires. Si une décision est répétée des milliers de fois, le coût des efforts nécessaires pour exercer à chaque fois un jugement aura tendance à devenir exorbitant, et il y aura sans doute beaucoup de bruit. Pour toutes les décisions répétées, il y a donc de vrais avantages à préférer la règle automatique au jugement *ad hoc*.

Les organisations intelligentes connaissent très bien les avantages et les inconvénients de ces deux manières de réguler les comportements. Elles

adoptent des règles, ou des directives qui s'en rapprochent, pour réduire le bruit (et le biais) dans les décisions répétitives. Et pour minimiser le coût des erreurs, elles sont prêtes à consacrer un temps et une attention considérables, en amont, à la définition des règles adéquates.

## *Mettre le bruit hors la loi ?*

Dans un grand nombre de situations, le bruit devrait être un scandale. Nous nous y sommes habitués, mais nous ne devrions pas le tolérer. Une solution simple est d'adopter des règles là où prévalent des principes qui laissent une large marge d'appréciation ; et nous savons maintenant quand cette solution simple a des chances d'être la bonne. Mais même quand il n'est pas souhaitable ou pas possible de fixer des règles, nous avons identifié toute une série de stratégies pour réduire le bruit.

Tout cela pose une question plus large : le système juridique devrait-il mettre le bruit hors la loi ? Il serait trop simple de répondre par l'affirmative, mais le droit devrait faire beaucoup plus qu'il ne le fait aujourd'hui pour contrôler le bruit.

Max Weber dénonçait ce qu'il appelait la « justice de cadi », qu'il définissait comme un ensemble de jugements *ad hoc* et informels, non contraints ou contrôlés par des règles générales. (Le cadi était le juge musulman dans l'ancien Empire ottoman, mais Weber considérait également comme pratiquant une « justice de cadi » les juges de paix anglais ou les jurys populaires français, dont les jugements étaient pris au cas par cas, en violation de l'autorité de la loi.) À ses yeux, la justice de cadi était intolérable : le bon juge devait « décider en fonction de règles formelles et "sans considération des personnes". Or, c'était tout le contraire : le cadi jugeait les personnes en fonction de leurs qualités

concrètes et de la situation concrète, ou selon l'équité et la convenance du résultat concret[4] ».

Cette approche, estimait Weber, « ne connaît pas de règles rationnelles pour la décision ». Weber, en somme, dénonçait le bruit intolérable produit par la justice de cadi. Il appelait de ses vœux l'avènement de la décision bureaucratique, ordonnée et encadrée en amont. (Rappelons-nous l'idée de justice bureaucratique, mentionnée plus haut sous la plume de Jerry Mashaw.) La spécialisation, la professionnalisation et la contrainte par la règle étaient à ses yeux l'étape ultime de l'évolution du droit.

Pourtant, aujourd'hui, bien après Weber, il est clair que la justice de cadi, ou qu'une justice qui lui ressemble, reste omniprésente. La question est de savoir ce que l'on peut y faire.

Nous n'irions pas jusqu'à dire que la réduction du bruit devrait faire partie de la Déclaration universelle des droits de l'homme. Mais dans certains cas, le bruit peut être considéré, au même titre que le biais, comme une violation des droits humains. Au minimum, les systèmes juridiques, dans le monde entier, devraient faire de plus grands efforts pour le contrôler.

Songeons à la peine pénale, à la demande d'asile, à la possibilité d'accéder à des études supérieures, à l'octroi de visas, de permis de construire ou de brevets. Est-il acceptable que ces décisions soient aussi largement aléatoires ? Ou rappelons-nous que les décisions de placement prises par les services de protection de l'enfance varient grandement en fonction des travailleurs sociaux auxquels elles sont confiées. Est-il acceptable que la vie et l'avenir de ces enfants et de leurs familles dépendent d'une loterie ? Nous voulons tous, bien sûr, que chacun de ces jugements soit fonction des particularités de chaque affaire. Mais qui peut trouver normal qu'elles soient en réalité fonction des particularités de chaque juge ?

Dans de nombreux cas, la variabilité de ces décisions apparaît, superficiellement ou après analyse, motivée par des biais. Qu'il s'agisse de biais cognitifs ou – surtout – de diverses formes de discrimination, presque tout le monde s'en indigne alors légitimement. L'injustice du biais est intolérable, et on en appelle volontiers à la loi pour la corriger. Les organisations, pour leur part, traquent impitoyablement les biais dans leurs décisions, qu'elles considèrent à juste titre comme une source d'erreurs coûteuses. Mais ni les citoyens, ni les dirigeants ne voient pour l'instant le bruit sous le même angle. Le biais nous indigne, le bruit nous laisse indifférents. Il est temps que cela change.

---

Dans d'innombrables champs de l'activité humaine, et en particulier partout où l'on applique des principes généraux plutôt que des règles précises, le niveau de bruit est élevé. Il impose des coûts élevés et produit des injustices. Encore les exemples dont nous avons parlé dans ce livre ne sont-ils que la partie émergée de l'iceberg, une toute petite partie d'un phénomène encore largement méconnu et mal mesuré. Le droit devrait faire beaucoup plus pour réduire ces coûts et combattre ces injustices.

### À propos des règles et des principes

« Nous nous contentons souvent de principes quand nous devrions adopter des règles – simplement parce qu'il est plus facile de se mettre d'accord sur des principes et parce que nous ne prêtons pas attention au bruit. »

« Des règles ou des principes ? Demandons-nous d'abord ce qui produira le moins d'erreurs. Puis demandons-nous quel effort sera nécessaire pour les produire et pour les appliquer. »

« La réduction du bruit ne devrait pas faire partie de la Déclaration universelle des droits de l'homme, du moins pas encore. Pourtant les systèmes juridiques devraient prendre des mesures fortes pour le réduire. »

SYNTHÈSE ET CONCLUSION

# Prendre le bruit au sérieux

Le bruit est la variabilité indésirable des jugements. Il y en a trop. Nos principaux objectifs dans ce livre ont été d'expliquer pourquoi il en est ainsi et de proposer des manières d'y remédier. Nous avons couvert beaucoup de sujets dans ce livre et nous aimerions, en guise de conclusion, en proposer une brève synthèse et donner une perspective.

## *Ce qu'est – et n'est pas – le jugement*

Jugement n'est pas synonyme de pensée. Le jugement, au sens où nous utilisons ici ce terme, est une activité bien plus précisément définie : une forme de mesure dont l'instrument est un esprit humain. Comme d'autres mesures, le jugement attribue à un objet une valeur choisie sur une échelle. Cette échelle peut être chiffrée ou non. « La tumeur de Mme Dupont est probablement bénigne » est un jugement, tout comme « l'économie du pays est très instable », « M. Durand serait le meilleur candidat pour le poste de directeur » et « la juste prime pour assurer ce risque est de 12 000 dollars ». Le processus de jugement consiste à intégrer de manière informelle différentes informations pour parvenir à une évaluation globale. Ce n'est

pas un calcul, et il ne suit pas des règles exactes. Un professeur formule un jugement quand il note une dissertation, pas quand il corrige un QCM.

Beaucoup de gens produisent des jugements dans l'exercice de leur profession. Ces professionnels du jugement, que nous appelons ici *juges* même quand ils n'ont rien à voir avec le système judiciaire, peuvent tout aussi bien être entraîneurs de football, cardiologues, ingénieurs, avocats, producteurs de cinéma, ou souscripteurs d'assurances. Le jugement professionnel est au centre de ce livre, à la fois parce qu'il a été largement étudié et parce que la qualité de ces jugements a un impact considérable sur nos vies. Par ailleurs, on peut raisonnablement supposer que les conclusions que nous tirons de l'étude des jugements professionnels peuvent s'appliquer aux jugements que nous produisons tous dans d'autres parties de nos vies.

Certains jugements sont *prédictifs*, et parmi ceux-ci, certains sont vérifiables : nous saurons un jour s'ils étaient justes ou non. C'est en général le cas pour les prévisions de court terme, comme celles portant sur les effets d'un médicament, l'évolution d'une pandémie ou l'issue d'une élection. Mais un grand nombre de jugements sont invérifiables : c'est le cas par exemple des prévisions à très long terme ou des réponses à des questions fictives. La qualité de ce genre de jugement ne peut pas être mesurée par référence à un résultat, mais uniquement en évaluant la qualité du processus de pensée qui l'a produit.

Par ailleurs, certains jugements ne sont pas du tout des prédictions, mais des *jugements évaluatifs.* À la différence des prédictions, les jugements évaluatifs ne peuvent pas se comparer à un résultat : la peine prononcée par un juge d'assises ou la note donnée à un vin dans un concours ne peuvent pas être considérées comme « justes » ou « fausses » par référence à des valeurs « vraies » objectivement définies.

Pourtant, cette distinction entre jugement prédictif et jugement évaluatif est rarement perçue. Les gens qui produisent des jugements se comportent généralement *comme si* une valeur vraie et objective existait et devait être

approchée au plus près, que cela soit ou non le cas. Ils agissent à la manière de tireurs visant une cible invisible : certains que la cible existe, même sans la voir, ils veulent s'en rapprocher le plus possible.

La croyance partagée qu'il existe une valeur juste, même quand elle ne peut pas être objectivement définie, fait partie intégrante de l'idée de jugement. Quand on parle d'une « question de jugement », on sous-entend qu'il est possible pour des individus raisonnables d'être en désaccord, mais aussi qu'il y a une limite à ce désaccord. Cette *anticipation de désaccord limité* caractérise les questions de jugement. Le jugement occupe ainsi une place intermédiaire entre les questions de calcul, où le désaccord n'est pas permis, et les questions de goût, où l'on ne s'attend pas particulièrement à être d'accord.

## *Les erreurs : biais et bruit*

Nous disons qu'il y a *biais* quand la plupart des erreurs d'un ensemble de jugements vont dans le même sens. Le biais est l'*erreur moyenne* ; il y a un biais, par exemple, quand des tireurs tirent toujours en bas et à gauche de la cible, quand des dirigeants d'entreprise font des prévisions de ventes trop optimistes année après année, ou quand une entreprise persiste à réinvestir de l'argent dans des projets qu'elle devrait abandonner.

Mais éliminer le biais d'un ensemble de jugements n'élimine pas toutes les erreurs. Les erreurs qui restent quand on a supprimé le biais sont celles qui ne sont *pas* partagées. Elles résultent de la variabilité indésirable des jugements, du manque de fiabilité de l'instrument de mesure que nous appliquons à la réalité. C'est ce que nous appelons le *bruit*. Le bruit est la variabilité de jugements qui devraient être identiques. Nous utilisons l'expression *bruit systémique* pour désigner le bruit que l'on observe dans les organisations où des professionnels présumés interchangeables doivent

prendre des décisions : des médecins dans un service d'urgence, des juges dans une cour d'assises, des experts en sinistres dans une compagnie d'assurances. Une grande partie de ce livre a traité du bruit systémique.

## *Mesurer le biais et le bruit*

L'*erreur quadratique moyenne* (*mean squared error*, ou MSE) est l'outil communément utilisé pour évaluer la justesse des mesures scientifiques depuis deux cents ans. La MSE a trois propriétés mathématiques qui sont importantes pour notre propos : elle est compatible avec l'intuition qui nous suggère que la moyenne d'un échantillon est la meilleure estimation de la moyenne de la population ; elle traite de la même manière les erreurs positives et les erreurs négatives ; et elle pénalise les grandes erreurs de façon plus importante que les petites.

Il pourrait sembler discutable d'utiliser la MSE pour mesurer l'erreur, car la MSE ne reflète pas le coût réel des erreurs de jugement. En particulier, les décisions professionnelles doivent prendre en compte le coût réel des erreurs, qui, dans le monde réel, est souvent asymétrique (c'est-à-dire que les erreurs sont plus graves dans un sens que dans l'autre). Toutefois, il est important de garder à l'esprit que toutes les décisions reposent sur des prédictions, et que ces prédictions doivent d'abord être justes. Si un ouragan menace la ville, les coûts associés à la sous-estimation ou à la surestimation de la menace ne sont évidemment pas les mêmes ; mais personne ne voudrait que ces coûts influencent la prévision que font les météorologistes quant à la trajectoire de la tempête. La MSE est la mesure d'erreur appropriée pour mesurer l'exactitude de ce genre de jugements prédictifs, et plus généralement de tous les jugements qui recherchent la justesse objective.

Quand l’erreur est mesurée par la MSE, le biais et le bruit sont des sources d’erreur indépendantes l’une de l’autre et qui s’additionnent. À l’évidence, le biais est toujours nuisible : le réduire améliorera nécessairement la justesse d’un jugement. Moins intuitif est le fait que le bruit est tout aussi néfaste, et que la réduction du bruit est, elle aussi, toujours une amélioration. Autrement dit, le degré optimal de variabilité dans les jugements est de zéro, même quand les jugements sont clairement biaisés. L’idéal, bien sûr, est de minimiser à la fois le biais et le bruit.

Le biais dans un ensemble de jugements vérifiables se définit par la différence entre la moyenne des jugements d’un cas donné et la valeur vraie correspondant à ce cas. Cette comparaison est impossible pour les jugements invérifiables. Par exemple, la valeur vraie d’une prime d’assurance pour un risque donné ne sera jamais connue ; pas plus qu’on ne pourra connaître la valeur vraie de la juste peine pour un crime donné. Dans ce type de cas, il est fréquent de faire l’hypothèse simplificatrice que les jugements ne sont pas biaisés, c’est-à-dire que la moyenne des jugements d’un grand nombre de juges est la meilleure estimation de la valeur vraie.

Le bruit, quant à lui, peut être évalué par un *audit de bruit*, c’est-à-dire par une expérience dans laquelle plusieurs professionnels produisent des jugements indépendants sur les mêmes affaires (réelles ou fictives). On peut bien sûr réaliser des audits de bruit sur des jugements vérifiables (par exemple en radiologie), mais on peut aussi le faire sur des jugements invérifiables (par exemple dans la justice pénale). Il est en effet possible de mesurer le bruit sans connaître la valeur vraie, tout comme l’on peut voir la dispersion d’un ensemble de tirs au dos d’une cible sans savoir où se trouve son centre. Un audit de bruit peut attirer l’attention sur des insuffisances de compétence et de formation. Il peut surtout quantifier le bruit systémique, par exemple quand des assureurs appartenant à une même équipe évaluent différemment les mêmes risques.

Du biais et du bruit, lequel est le plus problématique ? Cela dépend de la situation, mais on ne peut pas exclure *a priori* que ce soit le bruit. Pour que le biais et le bruit contribuent à égalité à l'erreur globale (MSE), il faut que la moyenne des erreurs (le biais) soit égale à l'écart-type des jugements (le bruit). Lorsque la distribution des jugements est « normale », c'est-à-dire qu'elle suit la courbe en cloche bien connue, cela se produit lorsque 84 % des jugements sont au-dessus (ou en dessous) de la valeur vraie. En d'autres termes, le biais est égal au bruit quand il est assez important pour que plus de cinq juges sur six se trompent dans le même sens. On peut raisonnablement supposer qu'un biais aussi important, dans un contexte professionnel, devrait souvent être détectable, et donc corrigé. Quand le biais est moins important, c'est le bruit qui est le principal contributeur à l'erreur globale.

## *Le bruit est problématique*

Dans certains jugements, la variabilité en tant que telle n'est pas un problème. Elle est même souvent bienvenue. La diversité des opinions est essentielle pour générer des idées et des options multiples. L'anticonformisme est essentiel à l'innovation. La pluralité des opinions chez les critiques de cinéma ou de littérature n'est pas une erreur, c'est un avantage. Les désaccords entre traders permettent à la bourse de fonctionner ; et les différences de stratégie entre start-up concurrentes permettent aux marchés de sélectionner la meilleure. Mais dans ce que nous appelons les questions de jugement, le bruit systémique est toujours un problème. Quand deux médecins font deux diagnostics différents du même patient, il y en a au moins un qui se trompe.

L'existence du bruit n'est pas une surprise. Ce qui en est une, c'est l'ampleur du bruit systémique et l'étendue des dégâts qu'il produit, qui

dépassent ce à quoi l'on s'attend généralement. Nous en avons donné des exemples dans de nombreux secteurs : l'entreprise, la médecine, la justice pénale, l'analyse des empreintes digitales, la prévision, l'évaluation du personnel et la politique. D'où notre conclusion : dès lors qu'il y a jugement, il y a bruit, et plus qu'on ne l'imagine.

Le rôle majeur du bruit dans l'erreur contredit une croyance largement répandue selon laquelle les erreurs aléatoires n'ont pas d'importance parce qu'elles s'annulent les unes les autres. Cette croyance est erronée. Si un grand nombre de tirs sont dispersés autour de la cible, cela n'a pas grand sens de dire que, « en moyenne », les tireurs ont mis dans le mille. Si une candidate à un emploi est mieux évaluée qu'elle ne le mérite et une autre moins bien, c'est la mauvaise personne qui sera embauchée. Si une police d'assurance est trop chère et une autre trop bon marché, les deux erreurs seront coûteuses pour la compagnie d'assurances : la première lui fera perdre des clients, la seconde lui fera perdre de l'argent. En somme, quand des jugements varient sans raison valable, nous pouvons être certains qu'il y a de l'erreur.

Même quand les jugements ne sont pas vérifiables, et quand il n'est donc pas possible de mesurer l'erreur, le bruit est néfaste. Il est injuste que des individus placés dans la même situation soient traités différemment, et un système qui produit des jugements professionnels incohérents perd vite sa crédibilité.

## *Les composantes du bruit*

La composante la plus intuitivement évidente du bruit systémique est le *bruit de niveau*. Certains magistrats sont généralement plus sévères, et d'autres plus cléments ; certains prévisionnistes sont généralement optimistes sur les perspectives du marché et d'autres pessimistes ; certains

médecins prescrivent plus d'antibiotiques quand d'autres s'en abstiennent, et ainsi de suite. Le bruit de niveau est la variabilité des jugements moyens produits par différents individus. Le bruit de niveau s'accroît quand les jugements utilisent des *échelles* ambiguës. Certains mots (par exemple *probablement*) ou certains chiffres (par exemple « 4 sur une échelle de 0 à 6 ») signifient en effet des choses différentes pour différentes personnes. Le bruit de niveau est une source majeure d'erreur et une cible de choix pour les actions de réduction du bruit.

Le bruit systémique comprend une autre composante, généralement plus importante. Indépendamment du niveau moyen des sentences qu'ils prononcent, deux magistrats qui ont à traiter des mêmes affaires ne seront pas nécessairement d'accord pour dire quels prévenus doivent être le plus lourdement sanctionnés. Leurs décisions produiront un classement différent des affaires. Nous appelons cette variabilité le *bruit de pattern* (en termes statistiques, on parle d'*interaction* entre le juge et l'affaire).

La principale source du bruit de pattern est stable : c'est la différence entre les réactions personnelles, idiosyncratiques, qu'évoque la même affaire chez différents individus. Certaines de ces différences reflètent des principes ou des valeurs qui guident les jugements de ces individus, consciemment ou non. Un juge peut ainsi se montrer particulièrement sévère avec les voleurs à la tire mais inhabituellement clément avec les auteurs d'infraction au Code de la route ; un autre peut montrer le pattern opposé. Les principes et valeurs sous-jacents peuvent être assez complexes, et le juge peut ne pas en être conscient : un juge peut ainsi, sans s'en rendre compte, être relativement clément avec les prévenus âgés accusés de vol à l'étalage. Enfin, une réaction très personnelle à un cas particulier peut aussi être stable. Si une prévenue ressemble à la fille du juge et lui inspire, sans qu'il en soit nécessairement conscient, une certaine clémence, le même effet aurait pu être constaté un autre jour.

On peut considérer ce *bruit de pattern stable* comme le reflet de la singularité des juges : les différentes affaires évoquent chez chaque individu un pattern de réactions caractéristique de cet individu, aussi singulier que sa personnalité. C'est en quelque sorte sa « personnalité de jugement ». Nous sommes habitués à considérer les différences subtiles entre personnalités comme intéressantes et enrichissantes, et dans beaucoup de contextes, bien sûr, elles le sont. Mais quand des professionnels opèrent dans un système qui recherche la cohérence, ces différences entre individus peuvent devenir problématiques. Dans les études que nous avons examinées, le bruit de pattern stable, qui est le produit de ces différences individuelles, est généralement la principale composante du bruit systémique.

Il faut toutefois souligner que les attitudes des individus vis-à-vis de cas particuliers ne sont pas parfaitement stables. Le bruit de pattern a aussi une composante transitoire, le *bruit occasionnel*. Le bruit occasionnel est évident quand un professionnel étudie le même cas en deux occasions différentes et produit des jugements différents : par exemple, quand un radiologue examine la même radiographie à différents moments de la semaine et fait des diagnostics différents, ou quand un expert en empreintes digitales réexamine une paire d'empreintes dont il a déjà jugé qu'elles se correspondaient, et conclut qu'elles ne se correspondent pas. Comme ces exemples le suggèrent, le bruit occasionnel est plus facile à mesurer dans les domaines où les juges ne se souviennent pas des cas qu'ils ont déjà examinés. Dans d'autres contextes, il existe un autre moyen de faire apparaître le bruit occasionnel : démontrer statistiquement l'effet sur les jugements d'un facteur contextuel qui ne devrait pas les affecter. Quand des juges sont plus cléments après une victoire de leur équipe de football favorite qu'après une défaite, par exemple, ou quand des médecins prescrivent davantage d'opioïdes en fin d'après-midi que le matin, on peut conclure à la présence de bruit occasionnel.

# *La psychologie du jugement et du bruit*

Dans les jugements prédictifs, les limites cognitives des juges ne sont pas, loin de là, la seule cause d'erreur. L'*ignorance objective* joue souvent un rôle plus important. L'ignorance objective porte sur des faits inconnaissables : combien de petits-enfants un enfant né hier aura-t-il à 70 ans ? Quel sera le numéro du billet de loterie gagnant la semaine prochaine ? Elle concerne aussi des faits qui peuvent théoriquement être connus, mais ne le sont pas du juge au moment de sa décision. Parce que nous sous-estimons notre ignorance objective (ainsi que nos biais), nous sommes généralement trop confiants dans la qualité de nos jugements prédictifs.

L'ignorance objective limite notre capacité à faire des prédictions exactes, souvent assez sévèrement. Cela ne nous empêche pas d'être généralement satisfaits de nos prédictions. Ce qui nous donne cette satisfaction n'est pas une validation externe, mais un *signal intérieur*, une gratification que nous nous autodécernons quand nous sommes parvenus à faire correspondre les faits à notre jugement et à les assembler en une histoire cohérente. Notre confiance subjective dans nos jugements n'a souvent rien à voir avec leur justesse objective.

La plupart des gens sont étonnés d'apprendre que la justesse de leurs jugements prédictifs n'est pas seulement faible, mais aussi inférieure à celle de formules assez simples. Des modèles de régression linéaire, construits à partir de données limitées, ou même des modèles « frugaux », à base de règles simples qui tiendraient sur le dos d'une enveloppe, font généralement mieux que les juges humains. L'avantage décisif des modèles et des règles, quels qu'ils soient, est d'être exempts de bruit. Nous vivons le jugement comme un processus subtil et complexe ; mais ce que nous prenons pour de la subtilité n'est le plus souvent que du bruit. C'est pourquoi une adhésion irréfléchie mais constante à des règles, même très simples, peut produire

des prédictions plus exactes qu'un jugement réfléchi mais inconstant. Ce résultat solidement établi demeure difficile à accepter, tant il contredit notre expérience subjective de l'acte de juger.

Les *biais psychologiques* sont bien sûr une source d'erreur systématique, c'est-à-dire de biais statistique. Mais, de manière moins évidente, ils sont aussi une source de bruit. Les biais psychologiques produisent du bruit quand les biais de tous les juges ne sont pas les mêmes, quand ils sont les mêmes mais présents à des degrés divers, ou quand leurs effets sont fonction de circonstances exogènes qui sont elles-mêmes variables. Pour prendre un exemple simple, si la moitié des recruteurs d'une société ont un biais qui défavorise les femmes, mais que l'autre moitié a un biais dans le sens opposé, il n'y aura pas, statistiquement, de biais sexiste dans les décisions de recrutement de l'entreprise. Il y aura pourtant beaucoup de bruit systémique, et un grand nombre d'erreurs de recrutement. Un autre exemple est l'effet disproportionné des premières impressions dans l'étude d'un dossier complexe : c'est un biais psychologique important, mais si l'ordre de présentation des éléments varie de façon aléatoire entre les dossiers, ce biais ne produira pas une erreur de direction prévisible, mais plutôt du bruit occasionnel.

Nous avons décrit le processus de jugement comme l'intégration informelle d'un ensemble d'informations pour produire une valeur choisie sur une échelle. L'élimination du bruit systémique nécessiterait donc que les juges maintiennent une parfaite uniformité dans leur utilisation des informations, dans le poids relatif qu'ils leur attribuent, et enfin dans leur utilisation de l'échelle. Même en laissant de côté les effets aléatoires du bruit occasionnel, ces conditions sont rarement remplies.

Le niveau d'accord est souvent assez élevé dans les jugements *unidimensionnels*. Des recruteurs différents seront souvent d'accord pour dire lequel de deux candidats est le plus charismatique ou le plus travailleur. Cette relative unanimité tient à une capacité que nous avons tous : nous

pouvons établir intuitivement des équivalences entre des jugements d'intensité sur des échelles différentes – par exemple, trouver la note à l'université qui nous semble correspondre avec une précocité dans la maîtrise de la lecture. Par conséquent, l'accord est souvent assez bon entre des juges qui évaluent une seule dimension sur une seule échelle. Il en va de même des jugements qui demandent d'intégrer un petit nombre d'informations convergeant toutes dans le même sens.

De grandes différences individuelles apparaissent en revanche quand un jugement exige de *pondérer de nombreuses informations contradictoires*. Face au même candidat, certains recruteurs vont attacher plus d'importance à son charisme ou à son intellect ; d'autres seront plus sensibles à son ardeur au travail ou à sa résilience dans les situations difficiles. Quand toutes les informations disponibles ne convergent pas pour bâtir une histoire cohérente, des juges différents vont inévitablement donner plus de poids à certaines informations et en ignorer ou en sous-pondérer d'autres. Cette variabilité dans le poids des différentes dimensions produira du bruit de pattern.

## *L'imperceptibilité du bruit*

Le bruit est un problème majeur. Pourtant, on n'en parle pratiquement jamais, certainement moins en tout cas qu'on ne parle du biais. Vous-même, avant de lire ce livre, n'y aviez sans doute jamais beaucoup songé. Étant donné son importance, l'inattention au bruit est un phénomène intéressant en soi.

Pour expliquer les erreurs de jugement, on invoque souvent les biais cognitifs, ainsi que d'autres distorsions de la pensée par l'émotion ou par l'intérêt. L'excès de confiance, l'ancrage, l'aversion à la perte et bien d'autres biais sont maintenant des explications courantes quand une

décision a mal tourné. Ces explications à base de biais sont satisfaisantes, parce que l'esprit humain a soif d'explications causales : quand quelque chose va mal, nous cherchons une explication, et généralement nous en trouvons une. Souvent, c'est un biais.

Le biais a une sorte de charisme explicatif dont le bruit est dépourvu. Si vous cherchez à expliquer, *a posteriori*, pourquoi une décision donnée était mauvaise, vous trouverez facilement du biais, mais vous ne penserez jamais au bruit. Pour appréhender le bruit, il faut adopter la *pensée statistique*, que nous avons également appelée *vision externe*, et qui consiste à considérer le cas qu'on étudie comme un élément d'un ensemble plus vaste. Malheureusement, la pensée statistique n'est pas naturelle : notre mode de raisonnement spontané est la *pensée causale*, qui construit des histoires cohérentes et logiques. La relative difficulté de la pensée statistique et le fait que nous y recourions si rarement contribuent à ce que le bruit soit plus difficile à remarquer que le biais.

À cette raison individuelle s'en ajoute une autre, organisationnelle : dans une entreprise ou une administration, les professionnels éprouvent rarement le besoin de se poser des questions sur le bruit dans leurs propres jugements ou dans ceux de leurs collègues. Passé leur formation initiale, ces professionnels sont souvent seuls face aux jugements qu'ils ont à faire. Les experts en empreintes digitales, les souscripteurs d'assurances expérimentés ou les employés aguerris du bureau des brevets n'interrompent pas souvent leur journée de travail pour se demander si leurs collègues pourraient ne pas être d'accord avec eux, et pourquoi. Il est encore plus rare qu'ils s'amusent à imaginer comment ils pourraient eux-mêmes être d'un avis différent si le même cas leur était présenté un autre jour.

Par conséquent, la plupart du temps, ces professionnels ont confiance dans la justesse de leurs propres jugements, et ils s'attendent à ce que leurs collègues soient d'accord avec eux. Mais ils n'ont en général l'occasion de vérifier ni cette justesse, ni cet accord : bien souvent, il est impossible de

comparer leur jugement à une « valeur vraie ». Au mieux, on pourra le soumettre à l'examen d'un autre professionnel considéré, à tort ou à raison, comme plus qualifié – ce que nous avons appelé un *expert d'estime*.

Ce n'est qu'exceptionnellement que des professionnels sont confrontés à la surprise d'un désaccord important entre eux. Quand cela se produit, ils trouvent généralement des raisons de considérer ce désaccord comme un cas isolé et de perpétuer ainsi leur *illusion d'accord,* c'est-à-dire leur croyance partagée dans l'absence de bruit systémique. Beaucoup de routines organisationnelles sont aussi conçues pour éviter les divergences entre experts, ou pour les masquer quand elles apparaissent. Ce réflexe est bien compréhensible : du point de vue d'une organisation, le bruit systémique a quelque chose de honteux.

## *Comment réduire le bruit (et le biais au passage)*

Certaines personnes produisent de meilleurs jugements que d'autres. Les meilleurs juges se caractérisent par leur compétence dans un domaine de jugement donné, par leur intelligence générale, mais aussi par un certain style cognitif, *activement ouvert d'esprit*. Ainsi, les bons juges feront moins d'erreurs. Néanmoins, même d'excellents juges, sur des problèmes de jugement complexe, ne seront pas toujours parfaitement d'accord. Les différences entre individus sont trop importantes pour cela. La variété infinie de leurs parcours, de leurs personnalités et de leurs expériences façonne leur personnalité de jugement. Cette infinie variété fait de chacun de ces juges une personne unique. Elle est aussi ce qui rend le bruit inévitable.

On peut lutter contre l'erreur en s'attaquant au biais. De façon générale, il existe deux manières répandues de corriger les biais dans les jugements : soit en corrigeant ces jugements après coup, soit en neutralisant un biais avant qu'il ne puisse les affecter (en changeant l'environnement de la décision ou en éduquant le décideur). Nous proposons une troisième option, qui s'applique tout particulièrement aux décisions prises en groupe : détecter les biais en temps réel en désignant un *observateur de la décision* chargé de repérer les signes que des biais sont présents (voir annexe B).

On peut également s'attaquer à l'erreur en réduisant le bruit, et la pratique que nous recommandons pour ce faire est l'*hygiène de la décision*. Comme l'hygiène sanitaire, la réduction du bruit dans le jugement est une forme de prévention à large spectre. Quand nous nous lavons les mains, nous empêchons des pathogènes de nous contaminer, sans savoir précisément lesquels. De même, l'hygiène de la décision permet d'éviter des erreurs sans que l'on sache de quelles erreurs il s'agit. Il y a sans doute moins de gloire et moins de satisfaction à prévenir l'apparition d'erreurs non identifiées qu'à corriger avec succès un biais bien précis. Et pourtant, le jeu en vaut largement la chandelle.

Pour s'engager dans la voie de la réduction du bruit, une organisation devrait toujours commencer par un audit de bruit (voir annexe A). L'audit est essentiel pour que l'organisation prenne le bruit au sérieux. Il permet également d'évaluer la part des différents types de bruit.

## *Six principes de réduction du bruit*

Nous avons évoqué les succès et les limites des efforts de réduction du bruit dans différents domaines. Nous récapitulons ici six principes qui sous-tendent l'hygiène de la décision, en soulignant comment ils répondent aux

mécanismes psychologiques qui génèrent le bruit et en rappelant les techniques d'hygiène de la décision qui illustrent ces principes.

*Principe 1. Puisque l'objectif du jugement est la justesse, ne pas rechercher l'expression individuelle*. Ce premier principe d'hygiène de la décision est une conséquence directe de la définition délibérément étroite du jugement que nous utilisons dans ce livre. Nous avons vu que le bruit de pattern stable est une composante importante du bruit systémique et qu'il résulte directement des différences entre individus. Des personnes différentes se forgent inévitablement des opinions différentes sur le même problème. Cette observation conduit à une conclusion impopulaire mais incontournable : une tâche de jugement ne sert pas à exprimer votre individualité.

Naturellement, l'individualité, la créativité et d'autres valeurs et préférences personnelles sont essentielles dans certaines phases de la réflexion et de la prise de décision. C'est le cas notamment quand on choisit ses objectifs, quand on imagine de nouvelles manières d'aborder un problème, et quand on génère des options. Mais quand vient le moment de produire un jugement sur ces options, et quand l'objectif est la justesse de ce jugement, toute expression d'individualité est une source de bruit. En d'autres termes, quand vous vous attendez à ce que les autres soient d'accord avec vous, vous devriez aussi essayer d'être d'accord avec eux, et vous demander ce que d'autres juges compétents penseraient s'ils étaient à votre place.

L'application radicale de ce principe est le remplacement du jugement humain par la règle ou l'algorithme. L'évaluation algorithmique élimine le bruit à coup sûr (c'est même la seule manière de l'éliminer entièrement). Les algorithmes sont déjà largement utilisés, et ils le seront de plus en plus. Il est cependant peu probable, dans un grand nombre de domaines, de les voir remplacer le jugement humain dans l'étape finale de décision – et c'est tant mieux.

Nous devons donc nous efforcer d'améliorer aussi le jugement humain. On peut pour cela s'inspirer de ce qui fait la force des algorithmes, en adoptant des approches qui rendent le jugement moins dépendant de l'individu qui juge et de ses caractéristiques personnelles. Nous avons constaté, par exemple, la puissance des directives : utilisées pour encadrer le pouvoir d'appréciation discrétionnaire des juges ou pour favoriser l'homogénéité des diagnostics des médecins, elles réduisent le bruit et améliorent les décisions.

*Principe 2. Adopter la vision externe et la pensée statistique*. Un juge adopte la vision externe quand il considère le cas auquel il s'intéresse non pas comme un cas unique, mais comme un élément d'un ensemble, un membre d'une *classe de référence* composée de tous les cas suffisamment similaires.

Ce mode de pensée n'est pas celui que nous adoptons par défaut : en général, nous nous focalisons sur le cas présent, et nous l'intégrons dans un récit causal qui vise à lui donner du sens. Or quand nous utilisons ce mode de raisonnement causal, le sens que nous donnons au cas présent et l'opinion que nous nous formons sur celui-ci s'enracinent inévitablement dans notre expérience personnelle et unique. D'autres juges voient le même cas différemment. Ces différences produisent du bruit de pattern.

C'est ce bruit de pattern que la vision externe aide à limiter, puisque des juges qui partagent la même classe de référence seront moins sujets au bruit dans leurs jugements. Qui plus est, la vision externe peut aider à corriger certains biais et donner des idées précieuses.

En permettant d'ancrer les prédictions sur les statistiques de cas similaires, la vision externe est la base de la pensée statistique, qui s'oppose à la pensée causale. La pensée statistique a deux conséquences importantes pour les jugements prédictifs. Elle conduit à recommander des prédictions plus modérées que ne le sont les prévisions intuitives (voir annexe C). Et elle aide les décideurs à éviter l'excès de confiance dans leurs jugements.

Prendre en considération de très nombreux cas passés les aide à prendre conscience de la grande variété et de la prédictibilité limitée de ces cas, et donc à mieux calibrer la confiance qu'ils peuvent avoir dans leurs jugements prédictifs. Il n'y a pas de honte à ne pas savoir prévoir l'imprévisible, mais c'est une faute de manquer d'humilité prédictive.

*Principe 3. Structurer les jugements en plusieurs tâches indépendantes*. Ce principe de décomposition vise à limiter les effets néfastes d'un mécanisme psychologique, l'*excès de cohérence* : nous avons tendance à déformer ou à ignorer les informations qui ne correspondent pas à un récit préexistant ou émergent ; et la justesse d'ensemble de nos jugements souffre quand des impressions qui portent sur des aspects distincts d'un cas se contaminent les unes les autres. La décomposition d'un problème de jugement en tâches plus petites et indépendantes les unes des autres limite ces contaminations croisées, à la manière d'un enquêteur qui veille à ce que les témoins d'un crime ne communiquent pas entre eux.

On trouve une application de ce principe dans la pratique de l'entretien d'embauche structuré, où les recruteurs évaluent le candidat sur chaque critère d'évaluation et notent leur conclusion avant de passer au critère suivant. On le retrouve également dans certaines directives en matière de diagnostic, comme le score d'Apgar.

Le principe de structuration est aussi le fondement du *protocole d'évaluations intermédiaires*. Ce protocole décompose un jugement complexe en plusieurs évaluations intermédiaires et veille à ce que chaque évaluation soit faite indépendamment des autres, par exemple en les confiant à des équipes différentes et en minimisant la communication entre elles.

*Principe 4. Résister aux intuitions prématurées*. Nous avons évoqué le *signal intérieur de fin de jugement*, qui permet aux décideurs de se sentir confiants dans leur jugement, même en l'absence d'éléments objectifs qui réduisent l'incertitude. Si les décideurs sont souvent réticents à utiliser des

directives, des algorithmes ou des règles qui restreignent leur marge d'appréciation, c'est en partie parce que ces méthodes ne leur donnent pas le même sentiment subjectif de confiance que le signal intérieur.

Les personnes qui décident, et qui doivent assumer les conséquences de leurs décisions, ont besoin d'être à l'aise avec leurs choix finaux. Le signal intérieur, ce sentiment gratifiant de confiance intuitive dans son propre choix, leur est nécessaire. L'important est qu'elles ne se l'accordent pas trop tôt. Choix intuitif n'est pas synonyme de jugement rapide : un choix intuitif sera bien meilleur s'il est informé par une analyse complète et équilibrée des éléments en présence. Le recours à l'intuition ne doit pas être prohibé, mais l'intuition doit être informée et disciplinée. Pour cela, elle doit donc être différée.

Le même principe inspire une autre recommandation pratique : *séquencer les informations*. Plus d'information n'est pas toujours une bonne chose, et de manière générale il n'est pas souhaitable de donner à des professionnels en situation de jugement des informations dont ils n'ont pas besoin, même si ces informations sont exactes. Les bonnes pratiques dans l'analyse des empreintes digitales, par exemple, sont désormais de ne donner aux experts aucune information contextuelle sur les suspects dont ils examinent les empreintes.

C'est également pour séquencer l'information et limiter la formation d'intuitions prématurées que le protocole d'évaluations intermédiaires recommande de contrôler l'ordre du jour des réunions de décision. Un ordre du jour efficace permet d'examiner séparément les différentes dimensions d'un problème et de différer la formation du jugement d'ensemble jusqu'à ce que toutes les dimensions aient été séparément évaluées.

*Principe 5. Obtenir plusieurs jugements indépendants avant de les agréger*. Dans la plupart des organisations, les habitudes de travail ne favorisent pas la formulation et le recueil de jugements réellement indépendants. C'est notamment le cas dans les réunions où les participants

s’influencent les uns les autres. À cause des effets de cascade et de la polarisation de groupe, ces discussions de groupe augmentent souvent le niveau de bruit. Le simple fait de recueillir les jugements individuels des participants *avant* la discussion révèle l’ampleur du bruit et facilite une résolution constructive des différences.

Faire la moyenne de plusieurs jugements indépendants garantit de réduire le bruit systémique (mais pas le biais). Deux ressorts distincts l’expliquent. D’abord, chaque jugement individuel peut être considéré comme un échantillon de la population de tous les jugements possibles, mais un échantillon de taille 1 : on améliore mécaniquement la précision d’une estimation en accroissant la taille de l’échantillon. Ensuite, l’avantage de la moyenne est plus grand encore quand les jugements émanent de juges aux compétences différentes et aux patterns de jugement complémentaires. La moyenne d’un groupe divers, donc sujet au bruit, peut s’avérer plus juste qu’un jugement unanime.

*Principe 6. Privilégier les jugements comparatifs et les échelles relatives*. Les jugements comparatifs sont moins sujets au bruit que les jugements absolus. Différentes personnes ont souvent une interprétation différente des échelles absolues, comme quand les mots « fréquent » ou « probable » sont diversement interprétés. Plus fondamentalement, notre capacité de faire des comparaisons deux à deux est bien supérieure à notre capacité de catégoriser sur une échelle. Les échelles qui font appel à des comparaisons sont donc moins sujettes au bruit que celles qui font appel à des jugements absolus. C’est ce qui fait la valeur des *échelles de cas*, définies par des cas de référence que les juges connaissent tous.

## *Les décisions singulières*

Les principes d'hygiène de la décision que nous venons d'évoquer démontrent leur efficacité dans les jugements récurrents où leurs effets peuvent être mesurés, mais ils sont aussi applicables aux décisions uniques que nous appelons des *décisions singulières*. L'existence du bruit dans les décisions singulières peut paraître contre-intuitive : par définition, on ne peut pas mesurer de variabilité quand on ne juge qu'une seule fois. Pourtant, le jugement singulier est sujet au bruit et aux erreurs qu'il occasionne. L'analogie de l'équipe de tireurs l'illustre : si nous ne voyons que le premier tireur en action, nous n'imaginons pas qu'il puisse y avoir du bruit. Pourtant, la dispersion des impacts deviendra visible quand ses coéquipiers auront tiré à leur tour. Le geste du premier tireur nous semble unique – jusqu'à ce qu'il soit répété. De même, la meilleure manière de concevoir le jugement singulier est de le traiter comme *un jugement récurrent qui ne se produit qu'une fois*. C'est pourquoi l'hygiène de la décision devrait l'améliorer lui aussi.

Pratiquer l'hygiène de la décision peut s'avérer ingrat. Le bruit est un ennemi invisible, et contre un ennemi invisible on ne remporte jamais qu'une victoire invisible. Cependant, comme l'hygiène sanitaire, l'hygiène de la décision est essentielle. Le patient qui récupère après une opération réussie attribue volontiers son salut à l'habileté du chirurgien, et il a raison, bien sûr. Mais si le chirurgien et tout le personnel présent en salle d'opération ne s'étaient pas lavé les mains, le patient n'aurait peut-être pas survécu. L'hygiène n'est pas glorieuse, mais elle est vitale.

## *Combien de bruit ?*

Bien sûr, les décideurs et les organisations ont bien d'autres préoccupations que la lutte contre le bruit. Le bruit peut ainsi être trop coûteux à réduire : si un lycée doit, pour éliminer le bruit dans ses notations, demander à cinq

professeurs de noter chaque devoir, cela ne semblera guère justifié. En pratique, un peu de bruit est souvent inévitable, notamment quand il est la conséquence d'une procédure qui accorde à chaque personne des garanties de traitement individualisé, qui ne traite pas les gens comme les rouages d'une machine, et qui laisse aux décideurs un sentiment d'autonomie d'action. Un peu de bruit pourrait même être utile, s'il permet au système de s'adapter au fil du temps – par exemple, quand le bruit qu'on observe reflète des valeurs et des objectifs qui ont changé, et suscite ainsi un débat nécessaire pour faire évoluer les pratiques ou les lois.

Une autre raison de tolérer du bruit est peut-être plus importante encore : certaines stratégies de réduction du bruit peuvent avoir des inconvénients inacceptables. Beaucoup de craintes entourant les algorithmes sont exagérées, mais certaines sont légitimes. Les algorithmes peuvent produire des erreurs stupides qu'un humain ne ferait jamais, et perdre ainsi en crédibilité, même s'ils permettent par ailleurs d'empêcher un grand nombre d'erreurs commises par les humains. Ils peuvent être biaisés par un défaut de conception ou par un entraînement sur des données inadaptées. Leur caractère impersonnel peut inspirer la méfiance. Quant aux pratiques d'hygiène de la décision, elles ont aussi leurs inconvénients : si elle est mal gérée, la réduction du bruit court le risque de bureaucratiser les décisions et de démotiver les professionnels, en leur donnant le sentiment que leur autonomie est rognée et leur compétence remise en question.

Ces arbitrages et ces risques méritent d'être pris pleinement en considération. Cependant, une objection (même légitime) à une stratégie de réduction du bruit n'implique pas de se résigner au bruit en général. On peut être contre l'agrégation des jugements dans une situation où l'on estime trop coûteux de solliciter des jugements multiples, et pourtant n'avoir rien contre la mise en place de directives. Et, dans d'autres situations, on peut faire le choix inverse.

En tout état de cause, une analyse coût-bénéfice s'impose : si les mesures qui permettent de réduire le bruit sont plus coûteuses que le bruit lui-même, la réduction du bruit est inopportune. Même sans aller jusque-là, l'analyse peut révéler un niveau de bruit optimal qui n'est pas de zéro. Le problème est qu'en l'absence d'audit de bruit, une organisation ne connaît pas la quantité de bruit dans ses jugements – et, rappelons-le une dernière fois, a généralement tendance à sous-estimer gravement cette quantité. Sans audit de bruit, invoquer la difficulté de la réduction du bruit n'est trop souvent qu'un prétexte pour continuer à se voiler la face en ne prenant pas la mesure du problème qu'il pose.

Le biais est source d'erreurs et d'injustice. Le bruit aussi ; et pourtant, nous ne nous en préoccupons guère. Une erreur de jugement peut certes sembler moins intolérable lorsqu'elle est aléatoire que lorsqu'elle peut être attribuée à une cause identifiée ; et pourtant une erreur imputable au bruit fait autant de dégâts que la même erreur causée par le biais. Le bruit est un problème important, sous-estimé, et traitable. Dans toutes les organisations, publiques et privées, celles et ceux qui ont à cœur d'améliorer la qualité de leurs décisions doivent prendre le bruit au sérieux.

ÉPILOGUE

# Moins de bruit

Imaginons à quoi ressemblerait un monde dans lequel les organisations auraient été repensées pour réduire le bruit. Les hôpitaux, les services de recrutement, les organismes de prévision économique, les agences gouvernementales, les compagnies d'assurances, les autorités de santé publique, la justice pénale, les cabinets d'avocats ou les universités seraient attentifs au problème du bruit et s'efforceraient de le réduire. L'audit de bruit serait entré dans les mœurs : on le répéterait chaque année pour mesurer les progrès accomplis.

Les dirigeants de ces organisations auraient accès, dans un nombre de domaines beaucoup plus important qu'aujourd'hui, à des algorithmes garantis sans biais, dont les résultats remplaceraient certains de leurs jugements et en assisteraient d'autres. L'habitude de décomposer les jugements complexes en évaluations intermédiaires plus simples serait une routine dans toutes les décisions importantes. D'autres techniques d'hygiène de la décision seraient largement connues et considérées comme des techniques de management ordinaires. Une vision externe serait intégrée de façon systématique dans les processus de décision pour favoriser l'utilisation de la pensée statistique.

Les réunions ne ressembleraient pas à celles que nous connaissons aujourd'hui : on y organiserait systématiquement l'expression de jugements indépendants, et la discussion y serait structurée autour d'ordres du jour définis à l'avance. Les désaccords ouverts seraient plus fréquents, mais aussi résolus de façon plus constructive. Chacun serait plus attentif à ce que le jugement des autres apporte à l'exactitude des décisions.

En somme, il y aurait moins de bruit dans le monde. Cela permettrait d'économiser beaucoup d'argent, d'améliorer la santé publique, de rendre la justice plus équitable, et de manière générale de prévenir de nombreuses erreurs évitables.

Notre objectif en écrivant ce livre était d'attirer l'attention sur cette opportunité. Ceux qui la saisiront rendront leurs organisations plus efficaces et le monde un peu plus juste. Nous espérons que vous en ferez partie.

GLOSSAIRE

# Jugement, erreur, biais, bruit

**Jugement, Juges.** Opération mentale (et résultat de cette opération mentale) consistant à associer à une impression une valeur sur une échelle, quantitative ou non.

- Les personnes qui produisent des jugements sont appelées *juges*.
- À la différence du calcul, où l'on n'attend pas de désaccord, et des opinions, où l'on tolère des désaccords importants, le jugement est caractérisé par une *anticipation de désaccord limité* entre juges compétents.

**Erreur.** Dans un jugement vérifiable, différence entre le jugement produit et la valeur vraie.

- L'erreur se compose de biais (erreur moyenne partagée avec d'autres juges) et d'une erreur de bruit (non partagée).

**Erreur quadratique moyenne (*mean squared error* ou MSE).** Dans un ensemble de jugements vérifiables sur des cas multiples, moyenne des carrés des erreurs (MSE).

- La variance observée des jugements se décompose en variance vraie (variance qui reflète la réalité des différences entre cas) et MSE (erreur ou variance indésirable).

**Biais (statistique).** Moyenne des erreurs.

- Dans le jugement d'un cas unique, moyenne des erreurs de jugement d'un ensemble de juges.
- Dans un ensemble de jugements sur des cas multiples, moyenne des erreurs de jugement.

**Biais (psychologique).** Erreur de jugement produite par l'utilisation d'une *heuristique* de jugement.

- Dans un système composé de juges multiples, un biais psychologique produit un biais statistique quand le biais est partagé par la majorité des juges, et du bruit systémique quand il ne l'est pas.

**Bruit.** Variabilité indésirable des jugements.

- Dans un jugement unique produit par un ensemble de juges, écart-type de leurs erreurs de jugement.

*Bruit systémique.* Variabilité indésirable des jugements dans un *système* où opèrent des juges multiples qui jugent des cas multiples.

- Le carré du bruit systémique est la part de la MSE qui n'est pas due au biais :

$$\text{MSE} = \text{biais}^2 + \text{bruit systémique}^2.$$

- Peut être mesuré par un *audit de bruit*.

- Souvent sous-estimé par les juges : c’est *l’illusion d’accord.*

**Bruit de niveau.** Part du bruit systémique due à l’effet principal des juges, c’est-à-dire aux différences dans les niveaux moyens de leurs jugements.

**Bruit de pattern.** Part du bruit systémique qui n’est pas due à l’effet principal des juges.

- bruit systémique$^2$ = bruit de niveau$^2$ + bruit de pattern$^2$

**Bruit occasionnel.** Part du bruit de pattern qui provient des variations intra-individuelles entre occasions de jugement ou de l’erreur aléatoire. Explicable notamment, mais jamais entièrement, par les influences perturbatrices de l’environnement de jugement.

**Bruit de pattern stable.** Part du bruit de pattern qui n’est pas due au bruit occasionnel, mais aux interactions stables entre juges et cas. Explicable notamment par les préférences, les valeurs et les personnalités des juges.

- bruit de pattern$^2$ = bruit occasionnel$^2$ + bruit de pattern stable$^2$

ANNEXE A

# Conduire un audit de bruit

Cette annexe est un guide pratique pour réaliser un audit de bruit. Il convient de le lire du point de vue d'un « consultant », qui peut être interne ou externe, chargé par une organisation d'examiner la qualité des jugements professionnels de ses employés en réalisant un audit de bruit dans un de ses services.

Comme son nom l'indique, l'audit vise à révéler la présence du bruit et son importance. Cependant, un audit bien conduit fournira également des informations utiles sur les biais, les points aveugles et les insuffisances dans la formation des employés et dans le contrôle de leur travail. Un audit réussi peut conduire à des changements dans l'organisation et les méthodes du service audité, mais aussi dans la doctrine qui guide les jugements des professionnels, dans la formation qu'ils reçoivent, dans les outils qu'ils utilisent pour étayer leurs jugements, et dans le contrôle auquel est soumis leur travail. Si le premier audit est un succès, il pourra être étendu à d'autres services de l'organisation.

L'audit de bruit demande un travail important et un grand souci du détail, car il faut s'attendre, si ses résultats font apparaître des insuffisances significatives, à ce que sa méthode soit critiquée. La procédure suivie et les cas utilisés doivent donc être conçus, dans les moindres détails, pour résister à ces critiques prévisibles. En particulier, les professionnels dont le

jugement est audité pourraient naturellement être les critiques les plus virulents (et les plus crédibles) de ses conclusions. Il est donc indispensable, pour désamorcer cette critique, qu'ils en soient, dès le départ, les principaux concepteurs.

En plus du consultant interne ou externe, l'audit de bruit fait intervenir quatre groupes de participants :

- *L'équipe projet*. L'équipe projet sera responsable de toutes les phases de l'étude. Si le consultant est interne, il sera le chef de cette équipe ; s'il est externe, une équipe projet interne travaillera en étroite liaison avec lui. L'objectif, dans les deux cas, est que l'organisation considère l'audit comme son projet, le consultant n'ayant qu'un rôle de soutien. Aux côtés du ou des consultants qui gèrent la collecte de données, l'analyse des résultats et la préparation du rapport final, l'équipe projet doit comprendre des experts du sujet, dont le rôle sera de construire des cas soumis aux juges. Ces experts doivent naturellement jouir d'une excellente crédibilité professionnelle aux yeux de leurs collègues.

- *Les clients*. L'audit de bruit n'est utile que s'il amène des changements importants. Il faut pour cela que les principaux dirigeants de l'organisation, « clients » du projet, soient impliqués dès le départ. On peut s'attendre, au départ, à ce que les clients soient sceptiques sur la prévalence du bruit, ou en tout cas à ce qu'ils le sous-estiment. Ce scepticisme initial est en réalité un avantage si les dirigeants sont ouverts d'esprit, curieux quant aux résultats de l'audit, et déterminés à remédier à la situation si l'audit démontre que le bruit est intolérable.

- *Les juges*. Les clients doivent choisir un ou plusieurs services à auditer. Le service sélectionné comprendra un nombre suffisant de « juges », définis comme des professionnels qui produisent des jugements et des décisions pour le compte de l'entreprise. Les juges doivent être

considérés comme plus ou moins interchangeables dans la pratique, c'est-à-dire qu'on considérerait comme normal et acceptable, en l'absence d'un des juges, qu'un autre traite la même affaire, et qu'on s'attendrait à ce qu'il arrive à un jugement similaire. C'était le cas par exemple pour les audits décrits en détail dans ce livre, qui portaient sur les juges fédéraux ainsi que sur les souscripteurs et les experts en sinistres dans une compagnie d'assurances. La tâche de jugement optimale pour un audit de bruit est un travail de jugement qui : (1) peut être réalisé à partir d'informations écrites ; et (2) se traduit par un jugement chiffré (en années de prison, en dollars, en probabilités, en notes…).

- *Le responsable de projet*. Un cadre de haut niveau doit veiller au bon déroulement du projet. Sa mission est de faciliter toutes les phases du projet, y compris la préparation du rapport final et la communication de ses conclusions à la direction de l'entreprise. Il n'est pas nécessaire qu'il soit un expert du sujet, mais son rang dans l'organisation est important, en pratique, pour surmonter les obstacles administratifs et pour signaler l'importance que l'organisation attache au projet.

## *La construction des cas*

Les experts du sujet qui font partie de l'équipe projet doivent avoir une expertise reconnue sur le sujet de l'audit (par exemple, la fixation de primes de risque ou l'évaluation du potentiel de certains investissements). Ils auront la responsabilité de développer les cas qui seront utilisés pour l'audit. Concevoir des cas qui constituent une simulation crédible des jugements réels est une tâche délicate, surtout compte tenu de l'examen critique auquel l'audit sera soumis s'il fait apparaître de sérieux problèmes.

L’équipe projet ne doit lancer l’audit qu’une fois certaine que, dans l’hypothèse où ses résultats indiqueraient un niveau de bruit élevé, les clients en concluront qu’il y a du bruit dans les jugements réels du service.

Il y a plusieurs façons de concevoir des cas crédibles. L’audit de bruit en matière de peine pénale (voir chapitre 1) résumait chaque cas par une liste schématique d’éléments pertinents et obtenait des évaluations sur seize cas en quatre-vingt-dix minutes. L’audit de bruit de la compagnie d’assurances (voir chapitre 2) se servait de résumés réalistes et détaillés de cas complexes. Ces deux audits apportèrent des éléments de preuve convaincants aux yeux de leurs « clients » respectifs en dépit du caractère schématique de ces cas, ou plutôt grâce à celui-ci : si les désaccords étaient fréquents et importants sur la base de cas très simplifiés, le bruit ne pouvait être que pire encore dans la réalité.

Un questionnaire doit accompagner chaque cas, afin de permettre d’analyser le raisonnement qui conduit chaque juge à produire un jugement sur ce cas. Ce questionnaire ne doit être rempli qu’une fois tous les cas examinés. Il doit comprendre :

- Des questions ouvertes sur les principaux facteurs qui ont conduit les juges à répondre comme ils l’ont fait.
- Une liste des éléments fournis sur chaque cas, demandant aux participants de noter l’importance de chaque élément.
- Des questions qui appellent une vision externe de la catégorie à laquelle le cas appartient. Par exemple, si les cas nécessitent une évaluation en dollars, on demandera aux juges d’estimer où chaque cas se situe dans la distribution de sa catégorie (par quintile ou par décile).

## *La réunion de lancement avec les « clients »*

Une fois les cas conçus, une réunion devra être organisée dans laquelle l'équipe projet présentera le plan de l'audit aux « clients », c'est-à-dire à la direction de l'organisation. L'objet de la réunion est de donner aux clients la possibilité de formuler des objections et des suggestions sur le plan de l'audit ; et d'obtenir d'eux l'engagement d'en accepter les résultats, y compris, le cas échéant, s'ils mettent en évidence un niveau inacceptable de bruit systémique. Faute de cet engagement, il serait inutile de passer à l'étape suivante. Si des objections sérieuses sont soulevées, l'équipe projet pourra améliorer le contenu des cas.

Une fois que les clients auront validé la conception de l'audit de bruit, l'équipe projet leur demandera de préciser les résultats auxquels ils s'attendent, en leur demandant par exemple :

- « À quel niveau de désaccord, exprimé en pourcentage, vous attendez-vous entre deux réponses choisies au hasard pour le même cas ? »
- « Quel est le niveau maximum de désaccord qui serait acceptable en termes d'efficacité opérationnelle ? »
- « À combien estimez-vous le coût d'une évaluation erronée dans un sens ou dans l'autre (par exemple, trop élevée ou trop basse de 15 %) ? »

Il est important de conserver la mémoire de ces réponses pour les comparer avec les résultats de l'audit une fois que ceux-ci seront dévoilés.

# *L’administration de l’étude*

Dès le départ, les responsables du service audité doivent être informés, en termes généraux, que leur service a été sélectionné pour en faire l’objet. Il est toutefois préférable que l’appellation *audit de bruit* ne soit pas utilisée pour qualifier le projet. Les mots *bruit* et *sujet au bruit* doivent être évités, surtout à propos d’individus. Une expression plus neutre, comme *étude du processus de prise de décision*, peut être employée.

Les responsables du service audité, avec l’assistance du responsable de projet et des membres de l’équipe projet, informeront les participants (« juges ») du lancement de l’étude. Son objectif peut être décrit en termes généraux, comme *« nous étudions la manière dont les [juges] arrivent à leurs conclusions »*.

Il est essentiel de donner aux professionnels qui participent à l’étude des assurances crédibles sur l’anonymat des réponses. Les réponses individuelles ne doivent être connues de personne au sein de l’organisation, même des membres de l’équipe projet. Si le projet est mené en interne, une entreprise extérieure (ou un outil informatique *ad hoc*) devra être utilisée pour anonymiser les données. Il est important également de souligner que le service choisi n’a été sélectionné que parce qu’il était représentatif des services qui produisent des jugements pour le compte de l’organisation, et que l’audit ne constitue pas une mise en cause de ses compétences. Pour garantir la crédibilité des résultats, tous les professionnels qualifiés du service devront participer à l’étude. L’étude proprement dite demande environ une demi-journée de travail, et le fait de lui allouer formellement ce temps contribuera à convaincre les participants de l’importance accordée à l’exercice.

Tous les juges doivent étudier leurs cas simultanément, mais il faut bien sûr les séparer physiquement les uns des autres et s’assurer qu’ils ne

communiquent pas entre eux pendant la durée de l’exercice. L’équipe projet sera disponible pour répondre à leurs questions.

## *Analyses et conclusions*

L’équipe projet réalisera les analyses statistiques des cas évalués par chaque participant, y compris la mesure de la quantité totale de bruit et de ses composantes, bruit de niveau et bruit de pattern. Si le contenu des cas le permet, l’analyse identifiera aussi des biais statistiques dans les réponses. L’équipe projet tentera également de comprendre les sources de la variabilité des jugements en examinant les réponses aux questionnaires dans lesquels les participants expliquent leur raisonnement. En s’intéressant notamment aux réponses extrêmes, l’équipe cherchera des patterns dans les données, éventuellement révélateurs de faiblesses dans la formation des employés, les procédures de l’organisation et les informations qu’elle fournit à ses employés.

Le consultant et l’équipe projet prépareront une synthèse des résultats qu’ils présenteront aux clients. Ils pourront également proposer des procédures pour améliorer les jugements et les décisions, en traitant les biais (si nécessaire) et en réduisant le bruit au moyen de techniques appropriées d’hygiène de la décision. Le développement, le test et la mise en œuvre de ces recommandations prendront naturellement plus longtemps que l’audit – généralement plusieurs mois.

Si l’audit du bruit de ce premier service a porté ses fruits, les dirigeants pourront décider de l’étendre à des jugements et décisions produits dans d’autres parties de l’organisation.

ANNEXE B

# Observer une décision

Cette annexe présente une check-list générique qui pourra être utilisée par un observateur de la décision (voir chapitre 19). Elle suit à peu près l'ordre chronologique de la discussion qui prépare une décision importante.

Les questions proposées sous chaque élément de la check-list apportent des clarifications supplémentaires. L'observateur de la décision doit se poser ces questions tout en observant le processus de décision.

Cette check-list n'est pas destinée à être utilisée telle quelle. Nous espérons plutôt qu'elle servira de point de départ aux observateurs de la décision, qui devront élaborer leur propre check-list.

## *Check-list d'observation des biais*

### 1. Approche du jugement

#### *1a. Substitution*

- « Le choix des informations utilisées par le groupe et leur discussion indiquent-ils qu'une question plus facile a été substituée à la question difficile qui leur était posée ? »

- « Le groupe a-t-il négligé un facteur important (ou semblé donner du poids à un facteur qui ne l'était pas) ? »

*1b. Absence de vision externe*

- « Le groupe a-t-il adopté une vision externe et essayé sérieusement d'utiliser cette vision externe pour formuler des jugements relatifs, plutôt qu'absolus ? »

*1c. Manque de diversité*

- « Y a-t-il une raison de penser que les membres du groupe partagent des biais qui pourraient se traduire par une corrélation de leurs erreurs ? À l'inverse, y a-t-il une expertise ou un point de vue importants qui ne sont pas représentés dans ce groupe ? »

2. PRÉJUGEMENTS ET CONCLUSION PRÉMATURÉE

*2a. Préjugements initiaux*

- « Les décideurs (ou certains d'entre eux) ont-ils plus à gagner d'une conclusion que d'une autre ? »
- « Quelqu'un était-il déjà convaincu d'une conclusion ? Y a-t-il des raisons de suspecter la présence de préjugés ? »
- « Les opposants ont-ils exprimé leurs opinions ? »
- « Y a-t-il un risque d'escalade de l'engagement dans une voie sans issue ? »

*2b. Conclusion prématurée ; excès de cohérence*

- « Y avait-il un biais accidentel dans le choix des éléments qui ont été discutés les premiers ? »
- « D'autres choix possibles ont-ils été pleinement pris en considération ? A-t-on activement recherché des éléments qui permettraient de les étayer ? »
- « Des données ou des opinions gênantes ont-elles été négligées ou étouffées ? »

## 3. Traitement de l'information

*3a. Disponibilité et visibilité*

- « Les participants exagèrent-ils la pertinence d'un événement ou d'une information en raison de son caractère récent ou spectaculaire, ou de sa résonance personnelle pour eux, même s'il n'a que peu de valeur objective ? »

*3b. Inattention à la qualité des informations*

- « Le jugement repose-t-il fortement sur des anecdotes, des histoires ou des analogies ? Si c'est le cas, des données plus complètes confirment-elles la représentativité de ces éléments ? »

*3c. Ancrage*

- « Des chiffres dont la pertinence ou l'exactitude sont douteuses jouent-ils un rôle important dans le jugement final ? »

*3d. Prédiction non régressive (par équivalence)*

- « Les participants ont-ils fait des extrapolations, des estimations ou des prévisions intuitives qui ne tiennent pas compte de la régression vers la moyenne ? »

## 4. Décision

*4a. Biais du planificateur*

- « Quand des prévisions ont été utilisées, a-t-on posé des questions sur leurs sources et leur validité ? La vision externe a-t-elle été utilisée pour mettre en perspective ces prévisions ? »
- « A-t-on utilisé des intervalles de confiance pour les chiffres incertains ? Sont-ils assez larges ? »

*4b. Aversion à la perte*

- « L'appétit pour le risque des personnes qui prennent les décisions est-il aligné sur celui de l'organisation ? Ces personnes sont-elles excessivement prudentes ? »

*4c. Biais pour le présent*

- « Les calculs (y compris le taux d’actualisation utilisé) reflètent-ils l’équilibre entre les priorités de court et de long terme de l’organisation ? »

ANNEXE C

# Corriger les prédictions intuitives

Les *prédictions par équivalence* sont les erreurs provoquées par le fait que nous établissons des équivalences intuitives entre des échelles d'intensité différentes (voir chapitre 14). Nous faisons une prédiction par équivalence quand nous nous comportons comme si les informations dont nous disposons prédisaient parfaitement (ou très fortement) le résultat que nous cherchons à prévoir.

Souvenons-nous de l'exemple de Julie, qui pouvait « lire couramment à l'âge de 4 ans ». La question était : quel est son GPA (sa note moyenne générale) à l'université ? Prédire un GPA de 3,8, c'était estimer que Julie, à 4 ans, était dans les meilleurs 10 % de son groupe d'âge pour la maîtrise de la lecture (mais pas dans le top 3-5 %). C'était surtout supposer, implicitement, que Julie, près de vingt ans plus tard, se situerait aussi autour du 90e percentile de sa classe en termes de GPA. Cette dernière supposition correspond à un GPA de 3,7 ou 3,8 – d'où la fréquence de ces réponses.

Ce raisonnement est statistiquement incorrect, parce qu'il surestime grossièrement la valeur diagnostique des informations qui se trouvent être disponibles sur Julie. Une enfant inhabituellement précoce à 4 ans ne réussit pas toujours de façon inhabituelle à l'université. Inversement, et fort

heureusement, un enfant qui a du mal à apprendre à lire n'est pas voué à tout jamais à faire partie des mauvais élèves !

Dans la plupart des cas, c'est même l'inverse qui se produit : une performance exceptionnelle se fait moins exceptionnelle avec le temps, et une performance médiocre tend à s'améliorer. Il est facile d'imaginer des facteurs sociaux, psychologiques ou politiques pour expliquer cette observation, mais aucune explication n'est en réalité nécessaire. Le phénomène est purement statistique. Les observations extrêmes dans un sens ou dans un autre ont tendance à se faire moins extrêmes avec le temps, simplement parce que la performance passée n'est pas corrélée de façon parfaite avec la performance future. Cette tendance est appelée *régression vers la moyenne* (d'où le terme *non régressives* pour les prédictions par équivalence, qui ne la prennent pas en compte).

On peut poser ce raisonnement en termes quantitatifs. Le jugement que vous avez produit sur Julie serait exact si l'âge de maîtrise de la lecture était un prédicteur parfait du GPA, c'est-à-dire s'il y avait une corrélation de 1 entre les deux variables. Ce n'est évidemment pas le cas.

Il existe une technique pour produire un jugement qui a de meilleures chances d'être exact. Elle n'est pas intuitive, et elle est rarement utilisée, même par des gens qui ont une formation en statistique. Nous la décrivons ici, en l'illustrant par l'exemple de Julie.

*1. Produire un jugement intuitif.* Votre intuition sur Julie, ou sur tout autre cas sur lequel vous avez des informations, n'est pas sans valeur. Votre pensée rapide – votre Système 1 – établit facilement une équivalence entre les informations dont vous disposez et l'échelle de GPA. Il vous donne ainsi un GPA pour Julie. Ce jugement intuitif serait la prédiction parfaite si les informations que vous avez étaient des prédicteurs parfaits du résultat. Notez-le.

*2. Chercher la moyenne.* Prenez maintenant un peu de recul et oubliez un moment ce que vous savez de Julie. Que diriez-vous de son GPA *si vous ne saviez absolument rien sur elle* ? La réponse est évidente : en l'absence de toute information, votre meilleure estimation du GPA de Julie serait le GPA moyen de sa classe – probablement autour de 3,2.

Considérer ainsi le cas de Julie, c'est appliquer le principe plus large dont nous avons parlé plus haut : la *vision externe*. Quand nous adoptons la vision externe, nous traitons le cas que nous examinons comme un élément de l'ensemble des cas comparables, et nous appliquons la pensée statistique à cet ensemble de cas. Rappelez-vous, par exemple, qu'adopter la vision externe sur le problème Gambardi nous avait conduits à nous interroger sur le « taux de base » de réussite d'un nouveau P-DG, c'est-à-dire sur la probabilité de succès moyenne d'un P-DG dont nous ne saurions rien (voir chapitre 4).

*3. Estimer la valeur prédictive des informations disponibles.* C'est l'étape difficile, où il faut que vous vous demandiez : « Quelle est la valeur prédictive réelle des informations dont je dispose ? »

La raison pour laquelle cette question est importante devrait maintenant être claire. Si tout ce que vous saviez de Julie était sa pointure de chaussures, vous n'accorderiez évidemment aucune importance à cette information, dont la corrélation avec le GPA est probablement de zéro, et vous vous en tiendriez logiquement, pour formuler votre prédiction, au GPA moyen. Si, à l'inverse, vous aviez la liste complète des notes que Julie a obtenues dans toutes les matières, ces informations seraient parfaitement prédictives de son GPA (qui n'est rien d'autre que leur moyenne). Il y a bien des nuances entre ces deux extrêmes. Si vous aviez des données sur les résultats scolaires de Julie au lycée, ces informations seraient beaucoup plus prédictives que l'âge où elle a su lire, mais moins que ses notes à l'université.

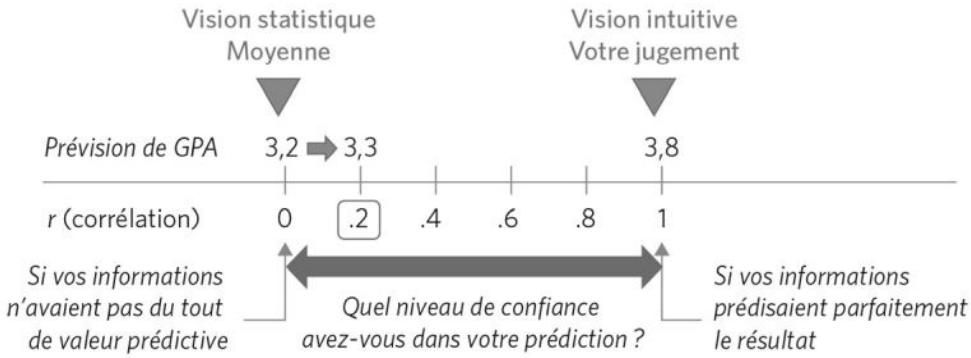



**Figure 19.** Corriger une prédiction intuitive pour tenir compte de la régression vers la moyenne.

Vous devez donc quantifier la valeur prédictive des données dont vous disposez, exprimée par leur *corrélation avec le résultat que vous voulez prédire*. Sauf exception, ce chiffre ne pourra être qu'une estimation au doigt mouillé, mais il sera utile quand même. Pour faire une estimation raisonnable, rappelez-vous les exemples que nous avons vus au chapitre 12. En sciences sociales, les corrélations de plus de 0,50 sont très rares. Un grand nombre de corrélations que nous jugeons significatives se situent autour de 0,20. Dans le cas de Julie, on peut supposer que la corrélation n'est pas supérieure à 0,20, et probablement inférieure.

*4. En partant de la vision externe, ajustez la prédiction dans la direction de votre jugement intuitif, dans une mesure qui correspond à la valeur prédictive des informations dont vous disposez.* La dernière étape, illustrée par la figure 19, est une simple combinaison arithmétique des trois chiffres que vous avez désormais produits : en partant de la moyenne, et en allant dans la direction de votre jugement intuitif, vous allez faire la fraction du chemin qui correspond à la corrélation que vous avez estimée.

Cette étape élargit simplement l'observation que nous venons de faire : si la corrélation était de 0, vous vous en tiendriez à la moyenne ; si elle était de 1, vous feriez à raison une prédiction par équivalence. Dans le cas de

Julie, si l'on suppose une corrélation de 0,20, la meilleure prédiction que vous puissiez faire de son GPA est celle qui s'écarte de 20 % de la moyenne de sa classe, dans le sens de l'estimation intuitive que vous a suggérée l'âge auquel elle a su lire. Ce calcul vous amène à une prédiction de 3,2 + ((3,8 – 3,2) × 0,20), soit environ 3,3.

Nous avons pris l'exemple de Julie, mais cette méthode peut tout aussi bien s'appliquer aux nombreux problèmes de jugement que nous avons évoqués dans ce livre. Imaginons, par exemple, un directeur des ventes qui recrute un nouveau vendeur et qui vient de rencontrer en entretien un candidat absolument exceptionnel. Sur la base de cette forte impression, il estime que le candidat devrait réaliser 1 million de dollars de chiffre d'affaires lors de sa première année de travail, ce qui correspond au double du montant atteint en moyenne par les nouveaux vendeurs pendant leur première année dans le poste. Comment pourrait-il corriger cette prévision par équivalence pour la rendre régressive ? La réponse dépend de la valeur diagnostique de l'entretien. Au vu des éléments que nous avons examinés sur la valeur prédictive d'un entretien d'embauche pour la performance au travail, nous savons qu'une corrélation de 0,40 est sans doute un grand maximum. Une estimation régressive des ventes de ce très impressionnant candidat serait donc, au mieux : 500 000 + ((1 million – 500 000) × 0,40) = 700 000 dollars.

Cette technique, répétons-le, n'est pas intuitive – et pour cause : elle sert à corriger l'intuition. En particulier, comme ce dernier exemple le montre, les prédictions corrigées seront toujours plus *modérées* que les prédictions intuitives. Elles ne seront jamais aussi extrêmes, et seront toujours plus proches, souvent *beaucoup* plus proches, de la moyenne. Si vous corrigiez vos prédictions, vous ne parieriez jamais que le joueur de tennis qui a gagné dix titres du Grand Chelem va en gagner dix autres. Vous ne prévoiriez jamais qu'une start-up qui a réalisé l'exploit de valoir 1 milliard de dollars

va devenir un mastodonte qui vaudra cent fois plus. Les prédictions corrigées ne parient jamais sur des exceptions ou des cas extrêmes.

La conséquence de ce principe est que les prédictions corrigées, quand elles sont évaluées *a posteriori*, feront inévitablement des erreurs extrêmement visibles : elles passeront toujours à côté des exceptions, qui sont souvent très présentes à l'esprit. Mais la prédiction ne se fait pas *a posteriori* ! Il faut garder à l'esprit que les exceptions sont, par définition, extrêmement rares. Il est beaucoup plus fréquent de faire l'erreur inverse, et de prédire que des exceptions vont se produire, ou que des performances exceptionnelles vont le rester. Quand nous faisons ce type de prédiction, en général, nous nous trompons, faute de prendre en compte la régression vers la moyenne. C'est pourquoi, chaque fois que l'objectif est de maximiser la justesse d'une prédiction, mesurée par l'erreur quadratique moyenne, les prédictions corrigées seront supérieures aux prédictions intuitives par équivalence.

# Notes

## Introduction

1. En utilisant des arcs et des flèches, plutôt que des carabines, le mathématicien suisse Daniel Bernoulli proposait la même analogie dans un essai de 1778 sur les problèmes d'estimation. Bernoulli, « Le choix le plus probable entre plusieurs observations contradictoires et la formation de l'induction la plus vraisemblable qui en résulte », *in* Kendall, « Studies in the history of probability and statistics : XI. Daniel Bernoulli on maximum likelihood », *Biometrika*, 1961, 48 (1-2), p. 3-18.

2. Joseph J. Doyle Jr, « Child protection and child outcomes : Measuring the effects of Foster care », *American Economic Review*, 2007, 95 (5), p. 1583-1610.

3. Stein Grimstad, Magne Jørgensen, « Inconsistency of expert judgment-based estimates of sotware development effort », *Journal of Systems and Software*, 2007, 80 (11), p. 1770-1777.

4. Andrew I. Schoenholtz, Jaya Ramji-Nogales, Philip G. Schrag, « Refugee roulette : Disparities in asylum adjudication », *Stanford Law Review*, 2007, 60 (2).

5. Mark A. Lemley, Bhaven Sampat, « Examiner characteristics and patent office outcomes », *Review of Economics and Statistics*, 2012, 94 (3), p. 817-827. Voir aussi Iain Cockburn, Samuel Kortum, Scott Stern, « Are all patent examiners equal ? The impact of examiner characteristics », *NBER Working Paper*, juin 2002, 8980, www.nber.org/papers/w8980 ; et Michael D. Frakes, Melissa F. Wasserman, « Is the time allocated to review patent applications inducing examiners to grant invalid patents ? Evidence from microlevel application data », *Review of Economics and Statistics*, 2017, 99 (3), p. 550-563.

# PARTIE I
# Détecter le bruit

# CHAPITRE 1
## Crime et châtiment

1. Marvin Frankel, « *Criminal Sentences : Law Without Order* », 25 Inst. for Sci. Info. Current Contents / Soc. & Behavioral Scis., *This Week's Citation Classic*, 1986, 14 (2A-6), http://www.garfield.library.upenn.edu/classics1986/A1986C697400001.pdf.

2. Marvin Frankel, *Criminal Sentences : Law Without Order*, New York, Hill and Wang, 1973, p. 5.

3. *Ibid.*, p. 103.

4. *Ibid.*, p. 5.

5. *Ibid.*, p. 11.

6. *Ibid.*, p. 114.

7. *Ibid.*, p. 115.

8. *Ibid.*, p. 119.

9. A. Partridge, William B. Eldridge, *A Report to the Judges of the Second Circuit. August 1974*, 1974, p. 9.

10. US Senate, *Comprehensive Crime Control Act of 1983 : Report of the Committee on the Judiciary, United States Senate, on S. 1762, Together with Additional and Minority Views*, Washington, DC, US Government Printing Office, 1983, rapport n° 98-225.

11. A. Partridge, William B. Eldridge, *A Report to the Judges of the Second Circuit. August 1974*, *op. cit.*, p. A11.

12. *Ibid.*, p. A9.

13. *Ibid.*, p. A5-A7.

14. William Austin, Thomas A. III Williams, « A survey of judges' responses to simulated legal cases : Research note on sentencing disparity », *The Journal of Criminal Law and Criminology*, 1977, 68 (2), p. 306.

15. John Bartolomeo *et al.*, « Sentence decisionmaking : The logic of sentence decisions and the extent and sources of sentence disparity », *The Journal of Criminal Law and Criminology*, 1981, 72 (2). Voir le chapitre 6 pour une présentation plus complète de cette étude. Voir aussi *Comprehensive Crime Control Act of 1983 : Report of the Committee on the Judiciary, United States Senate, on S. 1762, Together with Additional and Minority Views*, *op. cit.*, p. 44.

16. Shai Danziger, Jonathan Levav, Liora Avnaim-Pesso, « Extraneous factors in judicial decisions », *Proceedings of the National Academy of Sciences of the United States of America*, 2011, 108 (17), p. 6889-6892.

17. Ozkan Eren, Naci Mocan, « Emotional judges and unlucky juveniles », *American Economic Journal : Applied Economics*, 2018, 10 (3), p. 171-205.

18. Daniel L. Chen, M. Loecher, « Mood and the malleability of moral reasoning », *SSRN Electronic Journal*, 2019, p. 1-70.

19. Daniel L. Chen, Philippe Arnaud, « Mental accounting and social preferences : Judicial leniency on defendant birthdays », *Mimeo*, 2017, 614708, p. 1-33.

20. Anthony Heyes, Soodeh Saberian, « Temperature and decisions : Evidence from 207,000 court cases », *American Economic Journal : Applied Economics*, 2018, 11 (2), p. 238-265.

21. *Comprehensive Crime Control Act of 1983 : Report of the Committee on the Judiciary, United States Senate*, *op. cit.*, p. 38.

22. *Ibid*.

23. Cité par Jeffrey Rosen, « Breyer restraint », *New Republic*, 11 juillet 1994, p. 19-25.

24. *United States Sentencing Commission, Guidelines Manual*, 2018.

25. James M. Anderson *et al*., « Measuring interjudges sentencing disparity : Before and after the federal sentencing guidelines », *The Journal of Law and Economics*, 1999, 42 (S1), p. 271-308.

26. U.S. Sentencing Commission, *The Federal Sentencing Guidelines : A Report on the Operation of the Guidelines System and Short Term Impact on Disparity in Sentencing, Use of Incarceration, and Prosecutorial Discretion and Plea Bargaining*, 1991, vol. 1 et 2.

27. James M. Anderson *et al*., « Measuring interjudges sentencing disparity : Before and after the federal sentencing guidelines », art. cit.

28. Paul J. Hofer, Kevin R. Blackwell, R. Barry Ruback, « The effect of the Federal Sentencing Guidelines on inter-judge sentencing disparity », *The Journal of Criminal Law and Criminology*, 1999, 90 (1), p. 239-241.

29. Kate Stith, José Cabranes, *Fear of Judging : Sentencing Guidelines in the Federal Courts*, University of Chicago Press, 1998, p. 79.

30. 543 U.S. 220, 2005.

31. U.S. Sentencing Commission, *Results of Survey of United States District Judges, January 2010 through March 2010*, juin 2010, question 19, tableau 19.

32. Crystal Yang, « Have interjudge sentencing disparities increased in an advisory guidelines regime ? Evidence from Booker », *New York University Law Review*, 2014, 89, p. 1278-1334.

# CHAPITRE 2
## UN BRUIT SYSTÉMIQUE

1. Des cadres de la compagnie ont élaboré un descriptif détaillé de cas représentatifs, similaires aux risques et aux sinistres que les employés traitaient chaque jour : 6 pour les experts en sinistres dans le département incendies-risques divers, et 4 pour des souscripteurs spécialisés dans les risques financiers. Les employés consacrèrent une demi-journée à évaluer deux ou trois affaires chacun. Ils avaient l'instruction de travailler seuls et ne savaient pas que le but de l'étude était d'examiner la variabilité de leurs jugements. Au total, nous avons obtenu 86

jugements de 48 souscripteurs, et 113 jugements de 68 experts en sinistres. La méthode recommandée pour réaliser un audit de bruit est détaillée dans l'annexe A.

2. Dale W. Griffin, Lee Ross, « Subjective construal, social inference, and human misunderstanding », *Advances in Experimental Social Psychology*, 1991, 24, p. 319-359. Voir aussi Robert J. Robinson, Dacher Keltner, Andrew Ward, Lee Ross, « Actual versus assumed differences in construal : "Naive realism" in intergroup perception and conflict », *Journal of Personality and Social Psychology*, 68 (3), 1995, p. 404 *sq*. Voir également Lee Ross et Andrew Ward, « Naive realism in everyday life : Implications for social conflict and misunderstanding », *in* E. S. Reed, E. Turiel, T. Brown (dir.), *The Jean Piaget Symposium Series. Values and Knowledge*, Lawrence Erlbaum Associates, 1997.

# PARTIE II
# Votre esprit est un instrument de mesure

1. L'écart-type d'un ensemble de nombres vient d'une autre mesure statistique : la *variance*. Pour calculer la variance, il faut d'abord calculer les écarts par rapport à la moyenne de chacune des mesures, puis prendre le carré de chacun de ces écarts. La variance est la moyenne des carrés des écarts, et l'écart-type est la racine carrée de la variance.

## CHAPITRE 4
### Questions de jugement

1. R. T. Hodgson, « An examination of judge reliability at a major U.S. wine competition », *Journal of Wine Economics*, 2008, 3 (2), p. 105-113.

2. Certains chercheurs qui travaillent sur la prise de décision définissent la décision comme un choix entre différentes options et considèrent le jugement quantitatif comme une décision pour laquelle il existe un continuum de choix possibles. De ce point de vue, le jugement est un cas particulier de décision. Notre approche est différente : nous considérons la décision qui demande un choix entre des options comme issue d'un jugement évaluatif sous-jacent sur chacune de ces options, c'est-à-dire que nous considérons la décision comme un cas particulier de jugement.

## CHAPITRE 5
### Mesurer l'erreur

1. La méthode des moindres carrés a d'abord été publiée par Adrien-Marie Legendre en 1805. Gauss affirma qu'il l'avait déjà utilisée dix ans auparavant et l'associa plus tard à une théorie de l'erreur et à la courbe qui porte son nom. La querelle d'antériorité a fait l'objet d'âpres débats entre historiens, mais la plupart accordent la découverte à Gauss. Voir S. M. Stigler, « Gauss and the invention of least squares », *The Annals of Statistics*, 1981, 9, p. 465-474. S. M. Stigler, *The History of Statistics : The Measurement of Uncertainty before 1900*, Belknap Press of Harvard University Press, 1986.

2. Nous avons défini le bruit comme l'écart-type des erreurs ; le carré du bruit est donc la variance des erreurs. L'une des définitions de la variance est « la moyenne des carrés moins le carré de la moyenne ». La « moyenne des carrés » des erreurs est la MSE. Quant au « carré de la moyenne » des erreurs, puisque la moyenne des erreurs est le biais, c'est le carré du biais. On a donc : $\text{Bruit}^2 = \text{MSE} - \text{Biais}^2$. D'où l'équation de l'erreur.

3. Berkeley J. Dietvorst, Soaham Bharti, « People reject algorithms in uncertain decision domains because they have diminshing sensitivity to forecasting error », *Psychological Science*, 2020, 31 (10), p. 1302-1314.

# CHAPITRE 6
## L'ANALYSE DU BRUIT

1. Kevin Clancy, John Bartolomeo *et al.*, « Sentence decision making : The logic of sentence decisions and the extent and sources of sentence disparity », *The Journal of Criminal Law and Criminology*, 1981, 72 (2), p. 524-544. Voir également Inslaw, Inc. *et al.*, *Federal Sentencing – Towards a More Explicit Policy of Criminal Sanctions III-4*, 1981.

2. La sentence pouvait comprendre une peine d'incarcération, une durée de mise à l'épreuve et des amendes. Pour des raisons de simplicité, nous retenons ici le principal élément de la sentence – la durée de prison – et laissons les deux autres de côté.

3. Dans une situation de cas multiples et de juges multiples, on peut étendre l'équation de l'erreur que nous avons présentée dans le chapitre 5 pour y inclure un terme qui reflète cette variance. Spécifiquement, si nous définissons un « biais général » comme l'erreur moyenne de tous les cas, et si cette erreur n'est pas identique d'un cas à un autre, il y aura une *variance des biais des cas*. L'équation devient :

$$\text{MSE} = (\text{biais général})^2 + \text{variance des biais des cas} + \text{bruit systémique}^2.$$

4. Les chiffres mentionnés dans ce chapitre viennent de l'étude originale comme suit.
D'abord, les auteurs rapportent que l'effet principal du délit et du délinquant représente 45 % de la variance totale (Kevin Clancy, John Bartolomeo *et al.*, « Sentence decision making : The logic of sentence decisions and the extent and sources of sentence disparity », art. cit., tableau 6). Nous nous intéressons cependant plus largement à l'effet de chaque *affaire*, et donc à l'ensemble des éléments présentés aux juges – savoir, par exemple, si le prévenu avait un casier judiciaire ou si une arme a été utilisée. D'après notre définition, tous ces éléments sont légitimement pris en compte et relèvent de la « juste » variance de l'affaire, et non pas du bruit.

Nous avons donc réintégré les interactions entre les éléments de chaque affaire dans la variance de l'affaire (elles représentent 11 % de la variance totale, voir Kevin Clancy, John Bartolomeo *et al.*, « Sentence decision making : The logic of sentence decisions and the extent and sources of sentence disparity », art. cit., tableau 10). C'est pourquoi nous avons réévalué les parts dans la variance totale de la variance de l'affaire, de l'effet juge (bruit de niveau) et des interactions à, respectivement, 56 %, 21 % et 23 %. Le bruit est la somme des deux derniers chiffres, soit 44 % de la variance totale.
La variance des justes peines peut être calculée à partir du tableau qui fait la liste des peines moyennes pour chaque affaire (Kevin Clancy, John Bartolomeo *et al.*, « Sentence decision making : The logic of sentence decisions and the extent and sources of sentence disparity », art. cit., p. 89) : elle est de 15. Si cela représente 56 % de la variance totale, alors la variance totale est de 26,79 et la variance du bruit systémique de 11,79. La racine carrée de cette variance est l'écart-type d'une affaire représentative, soit 3,4 ans.
L'effet juge, ou le bruit de niveau, représente 21 % de la variance totale. La racine carrée de cette variance est l'écart-type attribuable au bruit de niveau du juge, soit 2,4 ans.

5. La valeur est la racine carrée de la moyenne des variances des peines pour les 16 affaires. Nous l'avons calculée comme indiqué dans la note ci-dessus.

6. La moyenne des valeurs absolues des différences entre deux observations dans une distribution normale d'écart-type 1 est de $\frac{2}{\sqrt{\pi}}$, soit 1,128.

7. L'hypothèse de l'additivité repose sur la supposition que la sévérité d'un juge ajoute une durée constante d'incarcération. Il est en réalité improbable que cela soit vrai : la sévérité du juge tendrait plutôt à ajouter une durée proportionnelle à la peine moyenne. Cette question a été ignorée dans le rapport original, qui ne donne aucun moyen d'en évaluer l'importance.

8. Kevin Clancy, John Bartolomeo *et al.*, « Sentence decision making : The logic of sentence decisions and the extent and sources of sentence disparity », art. cit., p. 23.

9. En utilisant l'équation : (bruit systémique)$^2$ = (bruit de niveau)$^2$ + (bruit de pattern)$^2$, et en considérant, d'après le tableau de la figure 9, que le bruit systémique est de 3,4 ans et le bruit de niveau de 2,4 ans, on peut calculer que le bruit de pattern est d'environ 2,4 ans. Le calcul est indicatif, et les valeurs véritables légèrement différentes à cause des erreurs d'arrondi.

# CHAPITRE 7
## Le bruit occasionnel

1. Voir http://www.iweblists.com/sports/basketball/FreeThrowPercent_c.html, consulté le 27 décembre 2020.

2. Voir https://www.basketball-reference.com/players/o/onealsh01.html, consulté le 27 décembre 2020.

3. R. T. Hodgson, « An examination of judge reliability at a major U.S. wine competition », *American Association of Wine Economists, 2008. Journal of Wine Economics*, 2008, 3 (2),

p. 105-113.

4. Stein Grimstad, Magne Jørgensen, « Inconsistency of expert judgment-based estimates of software development effort », *Journal of Systems and Software*, 2007, 80 (11), p. 1770-1777.

5. Robert H. Ashton, « A review and analysis of research on the test-retest reliability of professional judgment », *Journal of Behavioral Decision Making*, 2000, 13 (3), p. 277-294. Incidemment, l'auteur a observé qu'aucune des 41 études qu'il a examinées n'avait été *conçue* pour évaluer le bruit occasionnel : « Dans tous les cas, la mesure de la fiabilité était un sous-produit d'autres objectifs de recherche » (p. 279). Cela suggère que l'intérêt pour l'étude du bruit occasionnel est relativement récent.

6. Central Intelligence Agency, *The World Factbook*, Washington, 2020. Le chiffre cité inclut tous les aéroports ou aérodromes identifiables depuis les airs. La ou les pistes peuvent être bitumées ou non, et comprendre des installations fermées ou abandonnées.

7. Edward Vul, Harold Pashler, « Measuring the crowd within : Probabilistic representations within individuals », *Psychol. Sci.*, 2008, 19 (7), p. 645-648.

8. James Surowiecki, *The Wisdom of Crowds : Why the Many Are Smarter Than The Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*, Doubleday, 2004.

9. L'écart-type des jugements moyens (notre mesure du bruit) décroît en proportion de la racine carrée du nombre de jugements.

10. Stefan M. Herzog, Ralph Hertwig, « Think twice and then : Combining or choosing in dialectical bootstrapping ? », *Journal of Experimental Psychology : Learning Memory and Cognition*, 2014, 40 (1), p. 218-232.

11. Edward Vul, Harold Pashler, « Measuring the crowd within : Probabilistic representations within individuals », art. cit., p. 647.

12. Joseph P. Forgas, « Affective influences on interpersonal behavior », *Psychological Inquiry*, 2002, 13 (1), p. 1-28.

13. *Ibid.*, p. 10.

14. A. Filipowicz, A. Barsade, S. Melwani, « Understanding emotional transitions : The interpersonal consequences of changing emotions in negotiations », *Journal of Personality and Social Psychology*, 2011, 101 (3), p. 541-556.

15. Joseph P. Forgas, « She just doesn't look like a philosopher… ? Affective influences on the halo effect in impression formation », *European Journal of Social Psychology*, 2011, 41 (7), p. 812-817.

16. Gordon Pennycook *et al.*, « On the reception and detection of pseudo-profound bullshit », *Judgment and Decision Making*, 2015, 10 (6), p. 549-563 ; trad. fr. *De la réception et détection du baratin pseudo-profond, suivi de Baratin pour vous, transcendance pour moi, suivi de Ça reste du baratin*, Zones sensibles, 2016.

17. Harry Frankfurt, *On Bullshit*, Princeton University Press, 2005. Trad. fr. : *De l'art de dire des conneries*, Fayard, 2017.

18. Gordon Pennycook *et al.*, « On the reception and detection of pseudo-profound bullshit », art. cit., p. 549.

19. Joseph P. Forgas, « Happy believers and sad skeptics ? Affective influences on gullibility », *Current Directions in Psychological Science*, 2019, 28 (3), p. 306-313.

20. Joseph P. Forgas, « Mood effects on eyewitness memory : Affective influences on susceptibility to misinformation », *Journal of Experimental Social Psychology*, 2005, 41 (6), p. 574-588.

21. Piercarlo Valdesolo, David Desteno, « Manipulations of emotional context shape moral judgment », *Psychological Science*, 2006, 17 (6), p. 476-477.

22. Hannah T. Neprash, Michael L. Barnett, « Association of primary care clinic appointment time with opioid prescribing », *JAMA Network Open*, 2019, 2 (8) ; Lindsey M. Philpot *et al.*, « Time of day is associated with opioid prescribing for low back pain in primary care », *Journal of General Internal Medicine*, 2018, 33 (11), p. 1828-1830.

23. Jeffrey A. Linder, Jason N. Doctor, Mark W. Friedberg, Harry Reyes Nieva, Caroline Birks, Daniella Meeker, Craig R. Fox, « Time of day and the decision to prescribe antibiotics », *JAMA Intern Med.*, 2014, 174 (12), p. 2029-2031.

24. Rebecca H. Kim, Susan C. Day, Dylan S. Small, Christopher K. Snider, Charles A. L. Rareshide, Mitesh S. Patel, « Variations in influenza vaccination by clinic appointment time and an active choice intervention in the electronic health record to increase influenza vaccination », *JAMA Network Open*, 2018, 1 (5).

25. Joseph P. Forgas, Liz Goldenberg *et al.*, « Can bad weather improve your memory ? An unobtrusive field study of natural mood effects on real-life memory », *Journal of Experimental Social Psychology*, 2008, 45 (1), p. 254-257.

26. David Hirshleifer, Tyler Shumway, « Good day sunshine : Stock returns and the weather », *Journal of Finance*, 2003, LVIII (3), p. 1009-1032.

27. Uri Simonsohn, « Clouds make nerds look good : Field evidence of the impact of incidental factors on decision making », *The Journal of Behavioral Decision Making*, 2006, 20 (2), p. 143-152.

28. Daniel Chen *et al.*, « Decision making under the gambler's fallacy : Evidence from asylum judges, loan officers, and baseball umpires », *The Quarterly Journal of Economics*, 2016, 131 (3), p. 1181-1242.

29. Jaya Ramji-Nogales, Andrew I. Schoenholtz, Philip Schrag, « Refugee roulette : Disparities in asylum adjudication », *Stanford Law Review*, 60 (2).

30. Michael J. Kahana *et al.*, « The variability puzzle in human memory », *Journal of Experimental Psychology : Learning, Memory and Cognition*, 2018, 44 (12), p. 1857-1863.

# CHAPITRE 8
## LES GROUPES ET LE BRUIT

1. Matthew J. Salganik, Peter Sheridan Dodds, Duncan J. Watts, « Experimental study of inequality and unpredictability in an artificial cultural market », *Science*, 2006, 311, p. 854-856. Voir aussi : Matthew Salganik, Duncan Watts, « Leading the herd astray : An Experimental study of self-fulfilling prophecies in an artificial cultural market », *Social Psychology Quarterly*, 2008, 71, p. 338-355 ; Matthew Salganik, Duncan Watts, « Web-based experiments for the study of collective social dynamics in cultural markets », *Topics in Cognitive Science*, 2009, 1, p. 439-468.

2. Salganik et Watts, « Leading the herd astray », art. cit.

3. Michael Macy *et al.*, « Opinion cascades and the unpredictability of partisan polarization », *Science Advances*, 2019, 5 (8), p. 1-8. Voir aussi Helen Margetts *et al.*, *Political Turbulence*, Princeton University Press, 2015.

4. Michael Macy *et al.*, « Opinion cascades and the unpredictability of partisan polarization », art. cit.

5. Lev Muchnik *et al.*, « Social influence bias : A randomized experiment », 2013, *Science*, 341 (6146), p. 647-651.

6. Jan Lorenz *et al.*, « How social influence can undermine the wisdom of crowd effect », *Proceedings of the National Academy of Sciences*, 2011, 108 (22), p. 9020-9025.

7. Daniel Kahneman, David Schkade, Cass R. Sunstein, « Shared outrage and erratic awards : The psychology of punitive damages », *Journal of Risk and Uncertainty*, 1998, 16, p. 49-86.

8. David Schkade, Cass R. Sunstein, Daniel Kahneman, « Deliberating about dollars : The severity shift », *Columbia Law Review*, 2000, 100, p. 1139-1175.

# PARTIE III
# Le bruit dans les jugements prédictifs

1. Le pourcentage concordant (PC) est étroitement lié à la « concordance de Kendall », également appelée *W* de Kendall.

2. Kanwal Kamboj *et al.*, « A study on the correlation between foot length and height of an individual and to derive regression formulae to estimate the height from foot length of an individual », *International Journal of Research in Medical Sciences*, 2018, 6 (2), p. 528.

3. Le PC est calculé à partir de l'hypothèse que la distribution est normale bivariée. Les valeurs montrées ci-dessous sont des approximations fondées sur cette hypothèse. Nous remercions Julian Parris, qui a réalisé ce tableau.

# CHAPITRE 9

## JUGEMENTS ET MODÈLES

1. Martin C. Yu, Nathan R. Kuncel, « Pushing the limits for judgmental consistency : Comparing random weighting schemes with expert judgments », *Personnel Assessment and Decisions*, 2020, 6 (2). La corrélation de 0,15 obtenue par les experts est la moyenne non pondérée de trois échantillons étudiés pour un total de 847 cas. L'étude réelle diffère à plusieurs égards de cette description simplifiée.

2. Pour construire une moyenne pondérée, il est nécessaire que tous les indicateurs soient mesurés en unités comparables. Cette condition était remplie dans notre exemple introductif, où toutes les évaluations étaient faites sur une échelle de 0 à 10, mais ce n'est pas toujours le cas. Ainsi, les prédicteurs de performance future peuvent être l'évaluation faite par un interviewer sur une échelle de 0 à 10, le nombre d'années d'expérience pertinente et une note à un test d'aptitude. Le programme de régression multiple transforme tous les prédicteurs en notes normalisées (ou *valeurs centrées réduites*) avant de les combiner. Une note normalisée mesure la distance d'une observation à la moyenne d'une population, en utilisant l'écart-type pour unité. Par exemple, si la moyenne à un test d'aptitude est de 55 et que l'écart-type est de 8, une note normalisée de + 1,5 correspond à un résultat au test de 55 + 1,5*8 = 67. Une des conséquences importantes de la normalisation est de supprimer toute trace d'erreur dans la moyenne ou la variance des jugements des individus.

3. Une des caractéristiques importantes de la régression multiple est que le poids optimal de chaque prédicteur dépend des autres prédicteurs utilisés. Si un prédicteur est étroitement corrélé avec un autre déjà utilisé, son poids sera réduit pour éviter une forme de « double comptage ». Ce point sera développé à la fin du chapitre 21.

4. Robin M. Hogarth, Natalia Karelaia, « Heuristic and linear models of judgment : Matching rules and environments », *Psychological Review*, 2007, 114 (3), p. 734.

5. Un modèle largement utilisé dans ce contexte est le « modèle de la lentille », sur lequel cette discussion est fondée. Voir Kenneth R. Hammond, « Probabilistic functioning and the clinical method », *Psychological Review*, 1955, 62 (4), p. 255-262 ; Natalia Karelaia, Robin M. Hogarth, « Determinants of linear judgment : A meta-analysis of lens model studies », *Psychological Bulletin*, 2008, 134 (3), p. 404-426.

6. Paul Meehl, *Clinical Versus Statistical Prediction. A Theoretical Analysis and a Review of the Evidence*, University of Minnesota Press, 1954.

7. D. Kahneman, A. Tversky, « On the psychology of prediction », *Psychological Review*, 1973, 80, p. 237-251.

8. Paul Meehl, *Clinical Versus Statistical Prediction. A Theoretical Analysis and a Review of the Evidence*, Aronson, 1996, préface.

9. Paul E. Meehl, *in* Gardner E. Lindzey, William M. Runyan (dir.), *A History of Psychology in Autobiography*, American Psychological Association, 1989.

10. *Ibid.*, p. 362.

11. William M. Grove *et al.*, « Clinical versus mechanical prediction : A meta-analysis », *Psychological Assessment*, 2000, 12 (1), p. 19-30.

12. William Grove, Paul E. Meehl, « Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures : The Clinical-Statistical Controversy », *Psychology, Public Policy, and Law*, 1996, 2 (2), p. 293-323.

13. Lewis Goldberg, « Man versus model of man : A rationale, plus some evidence, for a method of improving on clinical inferences », *Psychological Bulletin*, 1970, 73 (6), p. 422-432.

14. Milton Friedman, Leonard J. Savage, « The utility analysis of choices involving risk », *The Journal of Political Economy*, 1948, 56 (4), p. 279-304.

15. Natalia Karelaia, Robin M. Hogarth, « Determinants of linear judgment : A meta-analysis of lens model studies », *Psychological Bulletin*, 2008, 134 (3), p. 404-426. Voir tableau 1, p. 411.

16. Nancy Wiggins, Eileen S. Kohen, « Man versus model of man revisited : The forecasting of graduate school success », *Journal of Personality and Social Psychology*, 1971, 19 (1), p. 100-106.

17. Natalia Karelaia, Robin M. Hogarth, « Determinants of linear judgment : A meta-analysis of lens model studies », art. cit.

18. La correction d'un coefficient de corrélation pour tenir compte de la fiabilité imparfaite d'un prédicteur est appelée *correction pour atténuation* ou *désatténuation*.

19. Martin C. Yu, Nathan Kuncel, « Pushing the limits for judgmental consistency : Comparing random weighting schemes with expert judgments », *Personnel Assessment and Decisions*, 2020, 6 (2).

20. Nous discuterons plus en détail des modèles équipondérés et des modèles aléatoires dans le prochain chapitre. Les poids sont limités à une série de petits nombres de même signe.

# CHAPITRE 10
## Le silence des formules

1. Robyn M. Dawes, Bernard Corrigan, « Linear models in decision making », *Psychological Bulletin*, 1974, 81 (2), p. 95-106. Dawes et Corrigan ont eux aussi proposé d'utiliser des poids aléatoires. L'étude de la prédiction des performances des cadres d'entreprise, décrite dans le chapitre 9, est une application de cette idée.

2. Jason Dana, « What makes improper linear models tick ? », *in* J. I. Krueger (dir.), *Rationality and Social Responsibility : Essays in Honor of Robyn M. Dawes*, Taylor & Francis, 2008, p. 73.

3. Jason Dana, Robyn M. Dawes, « The superiority of simple alternatives to regression for social sciences prediction », *Journal of Educational and Behavior Statistics*, 2004, 29, p. 317-331 ; Jason Dana, « What makes improper linear models tick ? », art. cit.

4. Le statisticien Howard Wainer développa ce raisonnement dans un article au titre mémorable : « It don't make no nevermind » (« Ça n'a tellement pas d'importance que ça ne

change rien à rien »). Howard Wainer, « Estimating coefficients in linear models : It don't make no nevermind », *Psychological Bulletin*, 1976, 83 (2), p. 213-217.

5. Jason Dana, « What makes improper linear models tick ? », art. cit., p. 72.

6. Il s'agit de convertir les données en notes normalisées, ou valeurs centrées réduites. Voir chapitre 9, note 2.

7. Martin C. Yu, Nathan R. Kuncel, « Pushing the limits for judgmental consistency : Comparing random weighting schemes with expert judgments », *Personnel Assessment and Decisions*, 2020, 6 (2), p. 1-10. Comme dans le chapitre précédent, la corrélation rapportée est la moyenne non pondérée des trois échantillons étudiés. La comparaison tient pour chacun des trois échantillons : la validité du jugement clinique expert était de 0,17, 0,16 et 0,13, et celle des modèles équipondérés de 0,19, 0,33 et 0,22 – respectivement.

8. Robyn M. Dawes, « The robust beauty of improper linear models in decision making », *American Psychology Review*, 1979, 34, p. 571-582.

9. Robyn M. Dawes, Bernard Corrigan, « Linear models in decision making », art. cit., p. 105.

10. Jongbin Jung, Conner Concannon, Ravi Shroff, Sharad Goel, Daniel G. Goldstein, « Simple rules to guide expert classifications », *Journal of the Royal Statistical Society, Statistics in Society*, 2020, 183, p. 771-800.

11. Julia Dressel, Hany Farid, « The accuracy, fairness, and limits of predicting recidivism », *Science Advances*, 2018, 4 (1), p. 1-6.

12. Ces deux exemples sont des modèles linéaires basés sur un nombre extrêmement réduit de variables (et, dans le cas du modèle de libération conditionnelle, sur une approximation des poids linéaires obtenus par une méthode d'arrondi qui transforme le modèle en calcul réalisable sans effort). Un autre type de modèle « impropre » est la *règle à variable unique*, qui ne considère qu'un seul prédicteur et ignore tous les autres. Voir Peter M. Todd, Gerd Gigerenzer, « Précis of simple heuristics that make us smart », *Behavioral and Brain Sciences*, 2000, 23 (5), p. 727-741.

13. P. Gendreau, T. Little, C. A. Goggin, « A meta-analysis of the predictors of adult offender recidivism : What works ! », *Criminology*, 1996, 34 (4), p. 575-607.

14. Dans ce contexte, la taille doit être comprise comme le rapport du nombre d'observations au nombre de prédicteurs. Dawes, dans « Robust beauty » (art. cit.), suggérait qu'il devait être de 15 à 20 pour 1 pour que la performance des modèles à pondération optimale dépasse, hors échantillon, celle des modèles équipondérés. Dans « Superiority of simple alternatives » (art. cit.), Dana et Dawes, utilisant beaucoup plus d'études de cas, mettaient la barre à un ratio de 100 pour 1.

15. J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, « Human decisions and machine predictions », *The Quarterly Journal of Economics*, 2018, 133 (1), p. 237-293.

16. L'algorithme a été entraîné sur un sous-ensemble de données, puis évalué sur sa capacité à prévoir les résultats sur un autre sous-ensemble, choisi au hasard.

17. J. Kleinberg *et al.*, « Human decisions and machine predictions », art. cit., p. 253.

18. Gregory Stoddard, Jens Ludwig, Sendhil Mullainathan, échanges de mails avec les auteurs, juin-juillet 2020.

19. B. Cowgill, « Bias and productivity in humans and algorithms : Theory and evidence from resume screening », article présenté à la Smith Entrepreneurship Research Conference, College Park, MD, 21 avril 2018.

20. William M. Grove, Paul E. Meehl, « Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures : The clinical-statistical controversy », *Psychology, Public Policy, Law*, 1996, 2 (2), p. 293-323.

21. Jennifer M. Logg, Julia A. Minson, Don A. Moore, « Algorithm appreciation : People prefer algorithmic to human judgment », *Organizational Behavior and Human Decision Processes*, 2019, 151, p. 90-103.

22. B. J. Dietvorst, J. P. Simmons, C. Massey, « Algorithm aversion : People erroneously avoid algorithms after seeing them err », *Journal of Experimental Psychology : General*, 2015, 144 (1), p. 114-126. Voir aussi : A. Prahl, L. Van Swol, « Understanding algorithm aversion : When is advice from automation discounted ? », *Journal of Forecasting*, 2017, 36 (6), p. 691-702.

23. M. T. Dzindolet, L. G. Pierce, H. P. Beck, L. A. Dawe, « The perceived utility of human and automated aids in a visual detection task », *Human Factors : The Journal of the Human Factors and Ergonomic Society*, 2002, 44 (1), p. 79-94. K. A. Hoff, M. Bashir, « Trust in automation : Integrating empirical evidence on factors that influence trust », *Human Factors*, 2015, 57, p. 407-434. P. Madhavan, D. A. Wiegmann, « Similarities and differences between human–human and human–automation trust : An integrative review », *Theoretical Issues in Ergonomics Science*, 2007, 8 (4), p. 277-301.

# CHAPITRE 11
## L'IGNORANCE OBJECTIVE

1. E. Dane, M. G. Pratt, « Exploring intuition and its role in managerial decision making », *Academy of Management Review*, 2007, 32 (1), p. 33-54. Cinla Akinci, Eugene Sadler-Smith, « Intuition in management research : A historical review », *International Journal of Management Reviews*, 2012, 14, p. 104-122. Gerard P. Hodgkinson *et al.*, « Intuition in organizations : Implications for strategic management », *Long Range Planning*, 2009, 42, p. 277-297.

2. Gerard P. Hodgkinson *et al.*, « Intuition in organization », art. cit., p. 279.

3. Nathan Kuncel *et al.*, « Mechanical versus clinical data combination in selection and admissions decisions : A meta-analysis «, *Journal of Applied Psychology*, 2013, 98 (6), p. 1060-1072. Voir aussi le chapitre 24 pour une discussion plus approfondie sur les décisions en matière de ressources humaines.

4. Don A. Moore, *Perfectly Confident : How to Calibrate Your Decisions Wisely*, HarperCollins, 2020.

5. Philip E. Tetlock, *Expert Political Judgment : How Good Is It ? How Can We Know ?*, Princeton University Press, p. 239 et 233.

6. William M. Grove *et al.*, « Clinical versus mechanical prediction : A meta-analysis », *Psychological Assessment*, 2000, 12 (1), p. 19-30.

7. Sendhil Mullainathan, Ziad Obermeyer, « Who is tested for heart attack and who should be : Predicting patient risk and physician error », *NBER Working Paper,* 2019, 26168.

8. Weston H. Agor, « The logic of intuition : How top executives make important decisions », O*rganizational Dynamics*, 1986, 14 (3), p. 5-18. Lisa A. Burke, Monica K. Miller, « Taking the mystery out of intuitive decision making », *Academy of Management Perspectives*, 1999, 13 (4), p. 91-99.

9. Poornima Madhavan, Douglas A. Wiegmann, « Effects of information source, pedigree, and reliability on operator interaction with decision support systems », *Human Factors : The Journal of the Human Factors and Ergonomics Society*, 2007, 49 (5), p. 773-785.

# CHAPITRE 12
## Dans la vallée des Choses Normales

1. Matthew J. Salganik *et al.*, « Measuring the predictability of life outcomes with a scientific mass collaboration », *Proceedings of the National Academy of Sciences*, 2020, 117 (15), p. 8398-8403.

2. Cet échantillon total comptait 4 242 familles, car certaines des familles de l'étude *Fragile Families* étaient exclues de l'analyse pour des raisons de protection de la vie privée.

3. Pour évaluer la justesse, les organisateurs du concours ont utilisé le même critère que celui que nous avons présenté dans la partie I : l'erreur quadratique moyenne, ou MSE. Pour faciliter les comparaisons, ils ont aussi comparé la MSE de chaque modèle à une stratégie de prévision « inutile », consistant à prédire pour chaque cas individuel la moyenne de l'échantillon d'entraînement. Par commodité, nous avons converti leurs résultats en coefficients de corrélation. La MSE et la corrélation sont liées par l'expression $r^2$ = (Var (Y) – MSE) / Var (Y), où Var (Y) est la variance du résultat et (Var (Y) – MSE) la variance des résultats prédits.

4. F. D. Richard *et al.*, « One hundred years of social psychology quantitatively described », *Review of General Psychology*, 2003, 7 (4), p. 331-363.

5. Gilles E. Gignac, Eva T. Szodorai, « Effect size guidelines for individual differences researchers », *Personality and Individual Differences*, 2016, 102, p. 74-78.

6. Une précision s'impose. Cette étude utilise à dessein un ensemble existant de données descriptives ; cet emsemble est très vaste, mais n'est pas spécifiquement conçu pour prévoir les résultats précis recherchés ici. C'est une différence importante avec les experts de l'étude de Tetlock, qui avaient toute discrétion d'utiliser les informations qu'ils jugeaient bonnes. Il pourrait être possible, par exemple, d'identifier des prédicteurs de l'expulsion d'une famille qui ne se trouvent pas dans ces données, mais qui auraient pu être collectés. Par conséquent, l'étude

ne mesure pas à quel point les expulsions (et d'autres événements) sont imprévisibles, mais à quel point elles sont imprévisibles *à partir de cet ensemble de* données utilisé par de nombreux sociologues.

7. Jake M. Hofman *et al*., « Prediction and explanation in social systems », *Science*, 2017, 355, p. 486-488 ; Duncan J. Watts *et al*., « Explanation, prediction, and causality : Three sides of the same coin ? », Center for Open Science, octobre 2018, p. 1-14, consultable sur https://osf.io/bgwjc.

8. Une autre distinction, très voisine, oppose la pensée « extensive » et la pensée « non extensive ». Amos Tversky et Daniel Kahneman, « Extensional versus intuitive reasoning : The conjunction fallacy in probability judgment », *Psychological Review*, 1983, 4, p. 293-315.

9. Daniel Kahneman, Dale T. Miller, « Norm theory. Comparing reality to its alternatives », *Psychological Review*, 1986, 93 (2), p. 136-153.

10. Baruch Fischhoff, « An early history of hindsight research », *Social Cognition*, 2007, 25 (1), p. 10-13 ; et « Hindsight is not equal to foresight : The effect of outcome knowledge on judgment under uncertainty », *Journal of Experimental Psychology : Human Perception and Performance*, 1975, 1 (3), p. 288.

11. Daniel Kahneman, *Thinking, Fast and Slow*, New York, Farrar, Straus and Giroux, 2011 ; trad. fr. *Système 1/Système 2, les deux vitesses de la pensée*, Flammarion, 2012.

# PARTIE IV
# D'où vient le bruit ?

## CHAPITRE 13
### HEURISTIQUES, BIAIS ET BRUIT

1. Daniel Kahneman, *Thinking, Fast and Slow*, New York, Farrar, Straus and Giroux, 2011 ; trad. fr. *Système 1/Système 2, les deux vitesses de la pensée*, Flammarion, 2012.

2. Une mise en garde s'impose. Les psychologues qui étudient les biais de jugement ne se satisfont pas de cinq participants par groupe, comme on le voit dans la figure 10, pour une bonne raison : comme les jugements sont sujets au bruit, les résultats de chaque groupe expérimental vont rarement se regrouper comme la figure 11 le suggère. Les gens varient dans leur susceptibilité à chaque biais et ne négligent pas *totalement* les variables pertinentes. Par exemple, avec un très grand nombre de participants, vous pourriez presque certainement confirmer que l'insensibilité à l'étendue est imparfaite : la probabilité moyenne attribuée au départ de Gambardi est très légèrement supérieure pour trois ans à ce qu'elle est pour deux ans.

La définition de l'insensibilité à l'étendue reste cependant appropriée parce que la différence est une fraction infime de ce qu'elle devrait être.

3. Daniel Kahneman *et al*. (dir.), *Judgment Under Uncertainty : Heuristics and Biases*, Cambridge University Press, 1982, chapitre 6 ; Daniel Kahneman, Amos Tversky, « On the psychology of prediction », *Psychological Review*, 1973, 80 (4), p. 237-251.

4. Voir, par exemple, Steven N. Kaplan, Bernadette A. Minton, « How has CEO turnover changed ? », *International Review of Finance*, 2012, 12 (1), p. 57-87. Voir aussi Dirk Jenter et Katharina Lewellen, « Performance-induced CEO turnover », Harvard Law School Forum on Corporate Governance, 2 septembre 2020.

5. J. W. Rinzler, *The Making of Star Wars : Return of the Jedi. The Definitive Story*, Del Rey, 2013, p. 64. Voir également Cass Sunstein, *The World According to Star Wars*, HarperCollins, 2016.

6. Nous soulignons ici le cas simple où le préjugement est préexistant. En réalité, un biais en faveur d'une certaine conclusion peut se développer à mesure que les preuves s'accumulent, en raison de la tendance à la cohérence et la simplicité. Au fur et à mesure que se dégage une conclusion provisoire, le biais de confirmation fait pencher les nouveaux éléments recueillis et leur interprétation dans le même sens.

7. Cette observation a été appelée le biais de croyance. Voir J. St. B. T. Evans, J. L. Barston, P. Pollard, « On the conflict between logic and belief in syllogistic reasoning », *Memory and Cognition*, 1983, 11 (3), p. 295-306.

8. Dan Ariely, George Loewenstein, Drazen Prelec, « "Coherent arbitrariness" : Stable demand curves without stable preferences », *The Quarterly Journal of Economics*, 2003, 118 (1), p. 73-105.

9. Adam D. Galinsky, T. Mussweiler, « First offers as anchors : The role of perspective-taking and negotiator focus », *Journal of Personality and Social Psychology*, 2001, 81 (4), p. 657-669.

10. On doit à Asch l'idée d'utiliser une série d'épithètes dans des ordres différents pour illustrer ce phénomène. Solomon E. Asch, « Forming impressions of personality », *Journal of Abnormal and Social Psychology*, 1946, 41 (3), p. 258-290.

11. Steven K. Dallas *et al*., « Don't count calorie labeling out : Calorie counts on the left side of menu items lead to lower calorie food choices », *Journal of Consumer Psychology*, 2019, 29 (1), p. 60-69.

# CHAPITRE 14
## ÉQUIVALENCES

1. S. S. Stevens, « On the operation known as judgment », *American Scientist*, 1966, 54 (4), p. 385-401. Notre emploi du terme « mise en équivalence » (*matching*) est plus large que celui de Stevens, qui se limitait aux échelles de rapports, sur lesquelles nous reviendrons dans le chapitre 15.

2. Cet exemple a d'abord été présenté dans Daniel Kahneman, *Thinking, Fast and Slow*, New York, Farrar, Straus and Giroux, 2011 ; trad. fr. *Système 1/Système 2, les deux vitesses de la pensée*, Flammarion, 2012.

3. Daniel Kahneman, Amos Tversky, « On the psychology of prediction », *Psychological Review*, 1973, 80, p. 237-251.

4. G. A. Miller, « The magical number seven, plus or minus two : Some limits on our capacity for processing information », *Psychological Review*, 1956, 63 (2), p. 63-97.

5. R. D. Goffin, J. M. Olson, « Is it all relative ? Comparative judgments and the possible improvement of self-ratings and ratings of others », *Perspectives on Psychological Science*, 2011, 6, p. 48-60.

# CHAPITRE 15
## ÉCHELLES

1. Daniel Kahneman, David Schkade, Cass Sunstein, « Shared outrage and erratic awards : The psychology of punitive damages », *Journal of Risk and Uncertainty*, 1998, 16, p. 49-86 ; et Cass Sunstein, Daniel Kahneman, David Schkade, « Assessing punitive damages (with notes on cognition and valuation in law) », *Yale Law Journal*, 1998, 107 (7), p. 2071-2153. Les coûts de la recherche ont été couverts par Exxon, mais l'entreprise n'a pas payé les chercheurs, ni contrôlé les données, ni eu accès aux résultats avant leur publication.

2. A. Keane, P. McKeown, *The Modern Law of Evidence*, Oxford University Press, 2014.

3. Andrew Mauboussin, Michael J. Mauboussin, « If you say something is ‘‘likely’’, how likely do people think it is ? », *Harvard Business Review*, 3 juillet 2018.

4. *BMW v. Gore*, 517 U.S. 559, 1996.

5. Pour une présentation du rôle de l'émotion dans le jugement moral, voir Jonathan Haidt, « The emotional dog and its rational tail : A social intuitionist approach to moral judgment », *Psychological Review*, 2001, 108 (4), p. 814-834. Voir également Joshua Greene, *Moral Tribes : Emotion, Reason, and the Gap Between Us and Them*, Penguin Books, 2014.

6. Compte tenu du volume important de bruit dans ces classements, vous pourriez être interloqué par la corrélation très forte (0,98) entre indignation et intention de sanctionner, qui étayait l'hypothèse de l'indignation. L'énigme s'évanouit quand on se rappelle que la corrélation a été calculée entre les *moyennes* des jugements. Quand on fait la moyenne de 100 jugements individuels, le bruit, qui est l'écart-type des jugements, est réduit d'un facteur 10. Le bruit cesse d'être un facteur quand de nombreux jugements sont agrégés. Voir chapitres 7 et 21.

7. S. S. Stevens, *Psychophysics : Introduction to Its Perceptual, Neural and Social Prospects*, John Wiley and Sons, 1975.

8. Dan Ariely, George Loewenstein, Drazen Prelec, « “Coherent arbitrariness” : Stable demand curves without stable preferences », *The Quarterly Journal of Economics*, 2003, 118 (1), p. 73-106.

9. Il y avait de nombreuses versions du questionnaire, puisque chaque individu ne jugeait que 10 scénarios sur les 28 existants. Nous avons réalisé l'analyse séparément pour chaque groupe de participants ayant considéré les mêmes 10 scénarios ; et les chiffres que nous rapportons ici sont une moyenne.

10. Une transformation en rangs implique une perte d'informations, car les écarts entre les jugements ne sont pas maintenus. Supposons qu'il n'y ait que trois affaires et que le juré recommande des dommages et intérêts de 10 millions, 2 millions et 1 million de dollars. Clairement, le juré entend exprimer une différence plus grande d'intention de sanctionner entre les deux premiers cas qu'entre le deuxième et le troisième. Une fois convertie en classements, cependant, la différence sera la même – une différence d'un rang seulement. Ce problème pourrait être résolu en convertissant les jugements en notes normalisées (valeurs centrées réduites) plutôt qu'en rangs (voir chapitre 9, note 2).

# CHAPITRE 16
## PATTERNS

1. R. Blake, N. K. Logothetis, « Visual competition », *Nature Reviews Neuroscience*, 2002, 3, p. 13-21 ; M. A. Gernsbacher, M. E. Faust, « The mechanism of suppression : A component of general comprehension skill », *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 1991, 17, p. 245-262 ; et M. C. Stites, K. D. Federmeier, « Subsequent to suppression : Downstream comprehension consequences of noun/verb ambiguity in natural reading », *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 2015, 41, p. 1497-1515.

2. D. A. Moore, D. Schatz, « The three faces of overconfidence », *Social and Personality Psychology Compass*, 2017, 11 (8), article e12331.

3. P. J. Lamberson, Scott Page, « Optimal forecasting groups », *Management Science*, 2012, 58 (4), p. 805-810. Nous remercions Scott Page d'avoir attiré notre attention sur cette source de bruit de pattern.

4. Le travail d'Allport et Odbert (1936) sur le vocabulaire anglais relatif à la personnalité est cité par Oliver P. John, Sanjay Srivastava, « The Big-Five trait taxonomy : History, measurement, and theoretical perspectives », *in* L. Pervin, Oliver P. John (dir.), *Handbook of Personality : Theory and Research*, Guilford, 1999, 2e éd.

5. Ian W. Eisenberg, Patrick G. Bissett, A. Zeynep Enkavi *et al*., « Uncovering the structure of self-regulation through data-driven ontology discovery », *Nature Communications*, 2019, 10, p. 2319.

6. Walter Mischel, « Toward an integrative science of the person », *Annual Review of Psychology*, 2004, 55, p. 1-22.

# CHAPITRE 17
## Les composantes du bruit

1. S'il n'y a pas de règle générale pour la décomposition du biais et du bruit, les proportions dans cette figure sont à peu près représentatives de certains des exemples, réels ou fictifs, que nous avons examinés. Dans ce graphique, en particulier, le biais et le bruit sont égaux (comme ils l'étaient dans les prévisions de ventes de GoodSell). Le carré du bruit de niveau représente 37 % du carré du biais systémique (comme dans l'étude sur les dommages punitifs). Le carré du bruit occasionnel est d'environ 35 % du carré du bruit de pattern.

2. Mark A. Lemley, Bhaven Sampat, « Examiner characteristics and patent office outcomes », *Review of Economics and Statistics*, 2012, 94 (3), p. 817-827. Voir aussi Iain Cockburn, Samuel Kortum, Scott Stern, « Are all patent examiners equal ? The impact of examiner characteristics », *NBER Working Paper*, juin 2002, 8980, www.nber.org/papers/w8980 ; et Michael D. Frakes, Melissa F. Wasserman, « Is the time allocated to review patent applications inducing examiners to grant invalid patents ? Evidence from microlevel application data », *Review of Economics and Statistics*, 2017, 99, p. 550-563.

3. Joseph J. Doyle Jr, « Child protection and child outcomes : Measuring the effects of foster care », *American Economic Review*, 2007, 95 (5), p. 1583-1610.

4. Andrew I. Schoenholtz, Jaya Ramji-Nogales, Philip G. Schrag, « Refugee roulette : Disparities in asylum adjudication », *Stanford Law Review*, 2007, 60 (2), p. 295-412.

5. Cette valeur est estimée à partir des calculs présentés dans le chapitre 6, où la variance d'interaction est de 23 % de la variance totale. D'après l'hypothèse que les peines sont normalement distribuées, la moyenne des valeurs absolues des différences entre deux observations sélectionnées de façon aléatoire est 1,128 SD.

6. J. E. Martinez, B. Labbree, S. Uddenberg, A. Todorov, « Meaningful "noise" : Comparative judgments contain stable idiosyncratic contributions » (non publié).

7. Du bruit – ou des différences de goût, si l'on estime qu'il ne s'agit pas ici à proprement parler d'une question de jugement. Il n'y a évidemment rien d'anormal à ce que les sujets de l'étude aient des préférences esthétiques différentes. L'intérêt de l'analyse n'est pas de mesurer ce total, mais de le décomposer entre sa composante stable et sa composante occasionnelle.

8. J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, « Human decisions and machine predictions », *The Quarterly Journal of Economics*, 2018, 133, p. 237-293.

9. Le modèle produisait pour chaque juge un classement des 141 833 affaires et un seuil au-delà duquel la liberté sous caution devait être accordée. Le bruit de niveau reflète la variabilité des seuils, et le bruit de pattern la variabilité de l'ordre des affaires.

10. Gregory Stoddard, Jens Ludwig, Sendhil Mullainathan, échanges de mails avec les auteurs, juin-juillet 2020.

11. Cette comparaison évoque celle de Montaigne, pour qui « il se trouve autant de différence de nous à nous mesmes, que de nous à autruy » (*Essais*, II, 1) ; mais notre conclusion, on le voit, est opposée.

12. Phil Rosenzweig, *Left Brain, Right Stuff : How Leaders Make Winning Decisions*, Public Affairs, 2014.

# PARTIE V
# Améliorer les jugements

# CHAPITRE 18
## De meilleurs juges pour de meilleurs jugements

1. Albert E. Mannes *et al.*, « The wisdom of select crowds », *Journal of Personality and Social Psychology*, 2014, 107 (2), p. 276-299 ; Jason Dana *et al.*, « The composition of optimally wise crowds », *Decision Analysis*, 2015, 12 (3), p. 130-143.

2. Briony D. Pulford, Andrew M. Colmna, Eike K. Buabang, Eva M. Krockow, « The persuasive power of knowledge : Testing the confidence heuristic », *Journal of Experimental Psychology : General*, 2018, 147 (10), p. 1431-1444.

3. Nathan R. Kuncel, Sarah A. Hezlett, « Fact and fiction in cognitive ability testing for admissions and hiring decisions », *Current Directions in Psychological Science*, 2010, 19 (6), p. 339-345.

4. *Ibid.*

5. Frank L. Schmidt, John Hunter, « General mental ability in the world of work : Occupational attainment and job performance », *Journal of Personality and Social Psychology*, 2004, 86 (1), p. 162.

6. Angela L. Duckworth, David Weir, Eli Tsukayama, David Kwok, « Who does well in life ? Conscientious adults excel in both objective and subjective success », *Frontiers in Psychology*, 2012, 3. Sur la « niaque » (*grit*), voir Angela L. Duckworth, Christopher Peterson, Michael D. Matthews, Dennis Kelly, « Grit : Perseverance and passion for long-term goals », *Journal of Personality and Social Psychology*, 2007, 92 (6), p. 1087-1101. Voir aussi : Angela Duckworth, *L'Art de la niaque*, J.-C. Lattès, 2017.

7. Richard E. Nisbett *et al.*, « Intelligence : New findings and theoretical developments », *American Psychologist*, 2012, 67 (2), p. 130-159.

8. Frank L. Schmidt, John Hunter, « General mental ability in the world of work : Occupational attainment and job performance », art. cit., p. 162.

9. Nathan R. Kuncel, Sarah A. Hezlett, « Fact and fiction in cognitive ability testing for admissions and hiring decisions », art. cit.

10. Ces corrélations proviennent de méta-analyses qui corrigent les corrélations observées pour tenir compte de l'erreur de mesure et de la restriction d'étendue (*range restriction*). Les chercheurs débattent encore de la question de savoir si ces corrections surestiment la valeur prédictive de la GMA. Mais comme ces questions méthodologiques s'appliquent aussi à d'autres prédicteurs, les experts considèrent en général que la GMA (avec les tests d'échantillon de travail, voir le chapitre 24) est le meilleur prédicteur actuel de la réussite au travail. Voir Nathan R. Kuncel, Sarah A. Hezlett, « Fact and fiction in cognitive ability testing for admissions and hiring decisions », art. cit.

11. Frank L. Schmidt, John Hunter, « General mental ability in the world of work : Occupational attainment and job performance », art. cit., p. 162.

12. David Lubinski, « Exceptional cognitive ability : The phenotype », *Behavior Genetics*, 2009, 39 (4), p. 350-358.

13. Jonathan Wai, « Investigating America's elite : Cognitive ability, education, and sex differences », *Intelligence*, 2013, 41 (4), p. 203-211.

14. Keela S. Thomson, Daniel M. Oppenheimer, « Investigating an alternate form of the cognitive reflection test », *Judgment and Decision Making*, 2016, 11 (1), p. 99-113.

15. Gordon Pennycook *et al.*, « Everyday consequences of analytic thinking », *Current Directions in Psychological Science*, 2015, 24 (6), p. 425-432.

16. Gordon Pennycook, David G. Rand, « Lazy, not biased : Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning », *Cognition*, 2018, 188, p. 39-50.

17. Nathaniel Barr *et al.*, « The brain in your pocket : Evidence that smartphones are used to supplant thinking », *Computers in Human Behavior*, 2015, 48, p. 473-480.

18. Niraj Patel, S. Glenn Baker, Laura D. Scherer, « Evaluating the cognitive reflection test as a measure of intuition/reflection, numeracy, and insight problem solving, and the implications for understanding real-world judgments and beliefs », *Journal of Experimental Psychology : General*, 2019, 148 (12), p. 2129-2153.

19. John T. Cacioppo, Richard E. Petty, « The need for cognition », *Journal of Personality and Social Psychology*, 1982, 42 (1), p. 116-131.

20. Stephen M. Smith, Irwin P. Levin, « Need for cognition and choice framing effects », *Journal of Behavioral Decision Making*, 1996, 9 (4), p. 283-290.

21. Judith E. Rosenbaum, Benjamin K. Johnson, « Who's afraid of spoilers ? Need for cognition, need for affect, and narrative selection and enjoyment », *Psychology of Popular Media Culture*, 2016, 5 (3), p. 273-289.

22. Wandi Bruine De Bruin *et al.*, « Individual differences in adult decision-making competence », *Journal of Personality and Social Psychology*, 2007, 92 (5), p. 938-956.

23. Heather A. Butler, « Halpern critical thinking assessment predicts real-world outcomes of critical thinking », *Applied Cognitive Psychology*, 2012, 26 (5), p. 721-729.

24. Uriel Haran, Ilana Ritov et Barbara Mellers, « The role of actively open-minded thinking in information acquisition, accuracy, and calibration », *Judgment and Decision Making*, 2013, 8

(3), p. 188-201.

25. *Ibid.*

26. J. Baron, « Why teach thinking ? An essay », *Applied Psychology : An International Review*, 1993, 42, p. 191-214 ; J. Baron, *Thinking and Deciding*, Cambridge University Press, 1994, 2e éd., p. 127-148.

# CHAPITRE 19

## Réduction des biais et hygiène de la décision

1. Pour une excellente synthèse, voir Jack B. Soll *et al.*, « A user's guide to debiasing », *in* Gideon Keren, George Wu (dir.), *The Wiley Blackwell Handbook of Judgment and Decision Making*, John Wiley and Sons, 2015, vol. 2, p. 684.

2. HM Treasury, *The Green Book : Central Government Guidance on Appraisal and Evaluation*, Londres, UK Crown, 2018, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/685903/The_Green_Book.pdf.

3. Richard H. Thaler, Cass R. Sunstein, *Nudge : Improving Decisions about Health, Wealth, and Happiness*, Yale University Press, 2008 ; trad. fr. *Nudge. La méthode douce pour inspirer la bonne décision*, Vuibert, 2010.

4. Ralph Hertwig, Till Grune-Yanoff, « Nudging and boosting : Steering or empowering good decisions », *Perspectives on Psychological Science*, 2017, 12 (6).

5. Geoffrey T. Fong *et al.*, « The effects of statistical training on thinking about everyday problems », *Cognitive Psychology*, 1986, 18 (3), p. 253-292.

6. Willem A. Wagenaar, Gideon B. Keren, « Does the expert know ? The reliability of predictions and confidence ratings of experts », *Intelligent Decision Support in Process Environments*, 1986, p. 87-103.

7. Carey K. Morewedge *et al.*, « Debiasing decisions : Improved decision making with a single training intervention », *Policy Insights from the Behavioral and Brain Sciences*, 2015, 2 (1), p. 129-140.

8. Anne-Laure Sellier *et al.*, « Debiasing training transfers to improve decision making in the field », *Psychological Science*, 2019, 30 (9), p. 1371-1379.

9. Emily Pronin *et al.*, « The bias blind spot : Perceptions of bias in self versus others », *Personality and Social Psychology Bulletin*, 2002, 28 (3), p. 369-381.

10. Daniel Kahneman, Dan Lovallo, Olivier Sibony, « Before you make that big decision… », *Harvard Business Review*, 2011, 89 (6), p. 50-60.

11. Atul Gawande, *The Checklist Manifesto : How to Get Things Right*, Metropolitan Books, 2010.

12. Office of Information and Regulatory Affairs, « Agency checklist : Regulatory impact analysis », non daté, www.whitehouse.gov/sites/whitehouse.gov/files/omb/inforeg/inforeg/regpol/RIA_Checklist.pdf.

13. Cette check-list est en partie adaptée de Daniel Kahneman *et al.*, « Before you make that big decision… », art. cit.

14. Voir Atul Gawande, *The Checklist Manifesto : How to Get Things Right*, *op. cit.*

# CHAPITRE 20

## POLICE SCIENTIFIQUE :

### LE SÉQUENCEMENT DE L'INFORMATION

1. R. Stacey, « A report on the erroneous fingerprint individualisation in the Madrid train bombing case », *Journal of Forensic Identification*, 2004, 54, p. 707-718.

2. Michael Specter, « Do fingerprints lie ? », *The New Yorker*, 27 mai 2002. C'est nous qui soulignons.

3. I. E. Dror, R. Rosenthal, « Meta-analytically quantifying the reliability and biasability of forensic experts », *Journal of Forensic Science*, 2008, 53, p. 900-903.

4. I. E. Dror, D. Charlton, A. E. Peron, « Contextual information renders experts vulnerable to making erroneous identifications », *Forensic Science International*, 2006, 156, p. 74-78.

5. I. E. Dror, D. Charlton, « Why experts make errors », *Journal of Forensic Identification*, 2006, 56, p. 600-616.

6. I. E. Dror, S. A. Cole, « The vision in "blind" justice : Expert perception, judgment, and visual cognition in forensic pattern recognition », *Psychonomic Bulletin and Review*, 2010, 17, p. 161-167, 165. Voir aussi I. E. Dror, « A hierarchy of expert performance (HEP) », *Journal of Applied Research in Memory and Cognition*, 2016, 5 (2), p. 1-6.

7. I. E. Dror *et al.*, « Cognitive issues in fingerprint analysis : Inter and intra-expert consistency and the effect of a "target" comparison », *Forensic Science International*, 2011, 208, p. 10-17.

8. B. T. Ulery, R. A. Hicklin, M. A. Roberts, J. A. Buscaglia, « Changes in latent fingerprint examiners' markup between analysis and comparison », *Forensic Science International*, 2015, 247, p. 54-61.

9. I. E. Dror, G. Hampikian, « Subjectivity and bias in forensic DNA mixture interpretation », *Science and Justice*, 2011, 51, p. 204-208.

10. M. J. Saks, D. M. Risinger, R. Rosenthal, W. C. Thompson, « Context effects in forensic science : A review and application of the science of science to crime laboratory practice in the United States », *Science Justice Journal of Forensic Science Society*, 2003, 43, p. 77-90.

11. President's Council of Advisors on Science and Technology (PCAST), *Report to the President. Forensic Science in Criminal Courts : Ensuring Scientific Validity of Feature-*

*Comparison Methods*, Washington, Executive Office of the President, PCAST, 2016.

12. R. Stacey, « A report on the erroneous fingerprint individualisation in the Madrid train bombing case », art. cit.

13. I. E. Dror, S. A. Cole, « The vision in “blind” justice : Expert perception, judgment, and visual cognition in forensic pattern recognition », art. cit.

14. I. E. Dror, « Biases in forensic experts », *Science*, 2018, 360, p. 243.

15. I. E. Dror, D. Charlton, « Why experts make errors », art. cit.

16. B. T. Ulery, R. A. Hicklin, J. A. Buscaglia, M. A. Roberts, « Repeatability and reproducibility of decisions by latent fingerprint examiners », *PLoS One*, 2012, 7 (3).

17. Innocence Project, « Overturning wrongful convictions involving misapplied forensics », *Misapplication of Forensic Science*, 2018, p. 1-7, www.innocenceproject.org/causes/misapplication-forensic-science. Voir aussi S. M. Kassin, I. E. Dror, J. Kukucka, L. Butt, « The forensic confirmation bias : Problems, perspectives, and proposed solutions », *Journal of Applied Research in Memory and Cognition*, 2013, 2 (1), p. 42-52.

18. PCAST, *Report to the President*, *op. cit*.

19. B. T. Ulery, R. A. Hicklin, J. Buscaglia, M. A. Roberts, « Accuracy and reliability of forensic latent fingerprint decisions », *Proceedings of the National Academy of Sciences*, 2011, 108, p. 7733-7738.

20. PCAST, *Report to the President*, *op. cit.*, p. 95. Les italiques sont dans l’original.

21. Igor Pacheco, Brian Cerchiai, Stephanie Stoiloff, *Miami-Dade Research Study for the reliability of the ACE-V Process : Accuracy and Precision in Latent Fingerprint Examinations*, rapport final, Miami-Dade Police Department Forensic Services Bureau, 2014, www.ncjrs.gov/pdffiles1/nij/grants/248534.pdf.

22. B. T. Ulery, R. A. Hicklin, M. A. Roberts, J. A. Buscaglia, « Factors associated with latent fingerprint exclusion determinations », *Forensic Science International*, 2017, 275, p. 65-75.

23. B. T. Ulery *et al*., « Accuracy and reliability of forensic latent fingerprint decisions », art. cit., p. 7735.

24. R. N. Haber, I. Haber, « Experimental results of fingerprint comparison validity and reliability : A review and critical analysis », *Science and Justice*, 2014, 54, p. 375-389.

25. I. E. Dror, « A hierarchy of expert performance (HEP) », art. cit., p. 3.

26. M. Leadbetter, lettre au rédacteur en chef, *Fingerprint Whorld*, 2007, 33, p. 231.

27. L. Butt, « The forensic confirmation bias : Problems, perspectives and proposed solutions – Commentary by a forensic examiner », *Journal of Applied Research in Memory and Cognition*, 2013, 2, p. 59-60. C’est nous qui soulignons.

28. R. Stacey, « A report on the erroneous fingerprint individualisation in the Madrid train bombing case », art. cit., p. 713. C’est nous qui soulignons.

29. J. Kukucka, S. M. Kassin, P. A. Zapf, I. E. Dror, « Cognitive bias and blindness : A global survey of forensic science examiners », *Journal of Applied Research in Memory and Cognition*, 2017, 6 (4).

30. I. E. Dror *et al*., lettre au rédacteur en chef : « Context management toolbox : A Linear Sequential Unmasking (LSU) Approach for minimizing cognitive bias in forensic decision making », *Journal of Forensic Science*, 2015, 60, p. 1111-1112.

# CHAPITRE 21

## PRÉVISION : L'AGRÉGATION ET LA SÉLECTION

1. Jeffrey A. Frankel, « Over-optimism in forecasts by official budget agencies and its implications », *NBER Working Paper*, 2011, 17239, www.nber.org/papers/w17239.

2. H. R. Arkes, « Overconfidence in judgmental forecasting », *in* Jon Scott Armstrong (dir.), *Principles of Forecasting : A Handbook for Researchers and Practitioners*, International Series in Operations Research and Management Science/Springer, 2001, vol. 30.

3. Itzhak Ben-David, John Graham, Campbell Harvey, « Managerial miscalibration », *The Quarterly Journal of Economics*, 2013, 128 (4), p. 1547-1584.

4. T. R. Stewart, « Improving reliability of judgmental forecasts », *in* Jon Scott Armstrong (dir.), *Principles of Forecasting*, *op. cit.*, p. 82.

5. Theodore W. Ruger, Pauline T. Kim, Andrew D. Martin, Kevin M. Quinn, « The Supreme Court Forecasting Project : Legal and political science approaches to predicting supreme court decision-making », *Columbia Law Review*, 2004, 104, p. 1150-1209.

6. Cass Sunstein, « Maximin », *Yale Journal of Regulation* (version préliminaire, 3 mai 2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3476250.

7. Pour de nombreux exemples, voir J. Scott Armstrong (dir.), *Principles of Forecasting*, *op. cit*.

8. Jon Scott Armstrong, « Combining Forecasts », *in* Jon Scott Armstrong (dir.), *Principles of Forecasting*, *op. cit.*, p. 417-439.

9. T. R. Stewart, « Improving reliability of judgmental forecasts », *in* Jon Scott Armstrong (dir.), *Principles of Forecasting*, *op. cit.*, p. 95. Voir aussi Véronique Genre *et al*., « Combining expert forecasts : Can anything beat the simple average ? », *International Journal of Forecasting*, 2013, 29 (1), p. 108-121.

10. Jon Scott Armstrong, « Combining Forecasts », art. cit.

11. Albert E. Mannes *et al*., « The wisdom of select crowds », *Journal of Personality and Social Psychology*, 2014, 107 (2), p. 276-299.

12. Justin Wolfers, Eric Zitzewitz, « Prediction markets », *Journal of Economic Perspectives*, 2004, 18, p. 107-126.

13. Cass R. Sunstein, Reid Hastie, *Wiser : Getting Beyond Groupthink to Make Groups Smarter*, Harvard Business Review Press, 2014.

14. Gene Rowe, George Wright, « The Delphi technique as a forecasting tool : Issues and analysis », *International Journal of Forecasting*, 1999, 15, p. 353-375. Voir aussi Dan Bang, Chris D. Frith, « Making better decisions in groups », *Royal Society Open Science*, 2017, 4 (8).

15. R. Hastie, « Review essay : Experimental evidence on group accuracy », *in* B. Grofman, G. Guillermo (dir.), *Information Pooling and Group Decision Making*, JAI Press, 1986, p. 129-157.

16. Andrew H. Van De Ven, Andre L. Delbecq, « The effectiveness of nominal, Delphi, and interacting group decision making processes », *Academy of Management Journal*, 2017, 17 (4).

17. Philip E. Tetlock, Dan Gardner, *Superforecasting : The Art and Science of Prediction*, Crown Publishers, 2015, p. 95.

18. *Ibid.*, p. 231.

19. *Ibid.*, p. 273.

20. Ville A. Satopää, Marat Salikhov, Philip E. Tetlock, Barbara Mellers, « Bias, information, noise : The BIN model of forecasting », 19 février 2020, p. 23, https://dx.doi.org/10.2139/ssrn.3540864.

21. *Ibid.*, p. 23.

22. *Ibid.*, p. 22.

23. *Ibid.*, p. 24.

24. Clintin P. Davis-Stober, David V. Budescu, Stephen B. Broomell, Jason Dana, « The composition of optimally wise crowds », *Decision Analysis*, 2015, 12 (3), p. 130-143.

25. P. J. Lamberson, Scott Page, « Optimal forecasting groups », *Management Science*, 2012, 58 (4), p. 805-810.

# CHAPITRE 22
## MÉDECINE : LES DIRECTIVES

1. Laura Horton *et al.*, « Development and assessment of inter- and intra-rater reliability of a novel ultrasound tool for scoring tendon and sheath disease : A pilot study », *Ultrasound*, 2016, 24 (3), p. 134, www.ncbi.nlm.nih.gov/pmc/articles/PMC5105362.

2. Laura C. Collins *et al.*, « Diagnostic agreement in the evaluation of image-guided breast core needle biopsies », *American Journal of Surgical Pathology*, 22004, 8, p. 126, https://journals.lww.com/ajsp/Abstract/2004/01000/Diagnostic_Agreement_in_the_Evaluation_of.15.aspx.

3. Julie L. Fierro *et al.*, « Variability in the diagnosis and treatment of group A streptococcal pharyngitis by primary care pediatricians », *Infection Control and Hospital Epidemiology*, 2014,

35 (S3), p. S79, www.jstor.org/stable/10.1086/677820.

4. Diabetes Tests, Centers for Disease Control and Prevention, https://www.cdc.gov/diabetes/basics/getting-tested.html, consulté le 15 janvier 2020.

5. Joseph D. Kronz *et al.*, « Mandatory second opinion surgical pathology at a large referral hospital », *Cancer*, 1999, 86 (11), p. 2426-2435.

6. La plupart du contenu peut se trouver en ligne. Voir aussi Dartmouth Medical School, *The Quality of Medical Care in the United States : A Report on the Medicare Program. The Dartmouth Atlas of Health Care 1999*, American Hospital Publishers, 1999.

7. Voir par exemple : OCDE, *Geographic Variations in Health Care : What Do We Know and What Can Be Done to Improve Health System Performance ?*, OECD Publishing, 2014, p. 137-169 ; Michael P. Hurley *et al.*, « Geographic variation in surgical outcomes and cost between the United States and Japan », *American Journal of Managed Care*, 2016, 22, p. 600, www.ajmc.com/journals/issue/2016/2016-vol22-n9/geographic-variation-in-surgical-outcomes-and-cost-between-the-united-states-and-japan ; et John Appleby, Veena Raleigh, Francesca Frosini, Gwyn Bevan, Haiyan Gao, Tom Lyscom, *Variations in Health Care : The Good, the Bad and the Inexplicable*, The King's Fund, 2011, www.kingsfund.org.uk/sites/default/files/Variations-in-health-care-good-bad-inexplicable-report-The-Kings-Fund-April-2011.pdf.

8. David C. Chan Jr *et al.*, « Selection with variation in diagnostic skill : Evidence from radiologists », *NBER Working Paper*, novembre 2019, 26467, www.nber.org/papers/w26467.

9. P. J. Robinson, « Radiology's Achilles' heel : Error and variation in the interpretation of the Röntgen image », *British Journal of Radiology*, 1997, 70, p. 1085, www.ncbi.nlm.nih.gov/pubmed/9536897. Pour une étude intéressante, voir Yusuke Tsugawa *et al.*, « Physician age and outcomes in elderly patients in hospital in the US : Observational study » (*BMJ*, 2017, 357, www.bmj.com/content/357/bmj.j1797), qui montre que plus le temps passe après la fin de la formation initiale des médecins, et plus leurs résultats se dégradent. Il s'ensuit qu'il y a un compromis entre le développement de l'expérience, qui vient des années de pratique, et la connaissance des données et des directives les plus récentes. L'étude montre que les médecins obtiennent les meilleurs résultats quand ils sont dans leurs premières années de pratique, et ont encore ces éléments en tête.

10. P. J. Robinson, « Radiology's Achilles' heel : Error and variation in the interpretation of the Röntgen image », art. cit.

11. Comme le coefficient de corrélation, le kappa peut être négatif, mais cela est rare en pratique. Voici une caractérisation de la signification de différentes statistiques kappa : « faible ($\kappa$ = 0,00 à 0,20), honnête ($\kappa$ = 0,21 à 0,40), modérée ($\kappa$ = 0,41 à 0,60), substantielle ($\kappa$ = 0,61 à 0,80) et presque parfaite ($\kappa$ > 0,80) ». Ron Wald, Chaim M. Bell, Rosane Nisenbaum, Samuel Perrone, Orfeas Liangos, Andreas Laupacis, Bertrand L. Jaber, « Interobserver reliability of urine sediment interpretation », *Clinical Journal of the American Society of Nephrology*, 2009, 4 (3), p. 567-571.

12. Howard R. Strasberg *et al.*, « Inter-rater agreement among physicians on the clinical significance of drug-drug interactions », *AMIA Annual Symposium Proceedings*, 2013, p. 1325,

www.ncbi.nlm.nih.gov/pmc/articles/PMC3900147.

13. R. Wald *et al.*, « Interobserver reliability of urine sediment interpretation », art. cit.

14. Juan P. Palazzo *et al.*, « Hyperplastic ductal and lobular lesions and carcinomas in situ of the breast : Reproducibility of current diagnostic criteria among community-and academic-based pathologists », *Breast Journal*, 2003, 4, p. 230.

15. Rohit K. Jain *et al.*, « Atypical ductal hyperplasia : Interobserver and intraobserver variability », *Modern Pathology*, 2011, 24, p. 917.

16. Alex C. Speciale *et al.*, « Observer variability in assessing lumbar spinal stenosis severity on magnetic resonance imaging and its relation to cross-sectional spinal canal area », *Spine*, 2002, 27, p. 1082.

17. Centers for Disease Control and Prevention, « Heart disease facts », www.cdc.gov/heartdisease/facts.htm, consulté le 16 juin 2020.

18. Timothy A. DeRouen *et al.*, « Variability in the analysis of coronary arteriograms », *Circulation*, 1977, 55, p. 324.

19. Olaf Buchweltz *et al.*, « Interobserver variability in the diagnosis of minimal and mild endometriosis », *European Journal of Obstetrics and Gynecology and Reproductive Biology*, 2005, 122, p. 213.

20. Jean-Pierre Zellweger *et al.*, « Intra-observer and overall agreement in the radiological assessment of tuberculosis », *International Journal of Tuberculosis & Lung Disease*, 2006, 10, p. 1123. Au sujet de la concordance interjuges « correcte », voir Yanina Balabanova *et al.*, « Variability in interpretation of chest radiographs among Russian clinicians and implications for screening programmes : Observational study », *BMJ*, 2005, 331, p. 379.

21. Shinsaku Sakurada *et al.*, « Inter-rater agreement in the assessment of abnormal chest X-ray findings for tuberculosis between two Asian countries », *BMC Infectious Diseases*, 2012, 12.

22. Evan R. Farmer *et al.*, « Discordance in the histopathologic diagnosis of melanoma and melanocytic nevi between expert pathologists », *Human Pathology*, 1996, 27, p. 528.

23. Alfred W. Kopf, M. Mintzis, R. S. Bart, « Diagnostic accuracy in malignant melanoma », *Archives of Dermatology*, 1975, 111, p. 1291.

24. Maria Miller, A. Bernard Ackerman, « How accurate are dermatologists in the diagnosis of melanoma ? Degree of accuracy and implications », *Archives of Dermatology*, 1992, 128, p. 559.

25. Craig A. Beam *et al.*, « Variability in the interpretation of screening mammograms by US radiologists », *Archives of Internal Medicine*, 1996, 156, p. 209.

26. P. J. Robinson *et al.*, « Variation between experienced observers in the interpretation of accident and emergency radiographs », *British Journal of Radiology*, 1999, 72, p. 323.

27. Katherine M. Detre *et al.*, « Observer agreement in evaluating coronary angiograms », *Circulation*, 1975, 52, p. 979.

28. Horton *et al.*, « Development and assessment of inter- and intra-rater reliability of a novel ultrasound tool for scoring tendon and sheath disease : A pilot study », art. cit. ; et Megan

Banky *et al.*, « Inter- and intra-rater variability of testing velocity when assessing lower limb spasticity », *Journal of Rehabilitation Medicine*, 2019, 51 (1), p. 54-60.

29. Esther Y. Hsiang *et al.*, « Association of primary care clinic appointment time with clinician ordering and patient completion of breast and colorectal cancer screening », *JAMA Network Open*, 2019, 51, https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2733171.

30. Hengchen Dai *et al.*, « The impact of time at work and time off from work on rule compliance : The case of hand hygiene in health care », *Journal of Applied Psychology*, 2015, 100, p. 846.

31. Ali S. Raja, « The HEART Score has substantial interrater reliability », *NEJM J. Watch*, 5 décembre 2018, www.jwatch.org/na47998/2018/12/05/heart-score-has-substantial-interrater-reliability [lecture de Colin A. Gershon *et al.*, « Inter-rater reliability of the HEART Score », *Academic Emergency Medicine*, 2019, 26, p. 552].

32. Jean-Pierre Zellweger *et al.*, « Intra-observer and overall agreement in the radiological assessment of tuberculosis », *International Journal of Tuberculosis and Lung Disease*, 2006, 10, p. 1123 ; Ibrahim Abubakar *et al.*, « Diagnostic accuracy of digital chest radiography for pulmonary tuberculosis in a UK urban population », *European Respiratory Journal*, 2010, 35, p. 689.

33. Michael L. Barnett *et al.*, « Comparative accuracy of diagnosis by collective intelligence of multiple physicians vs individual physicians », *JAMA Network Open*, 2019, 2, https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2726709 ; Kimberly H. Allison *et al.*, « Understanding diagnostic variability in breast pathology : Lessons learned from an expert consensus review panel », *Histopathology*, 2014, 65, p. 240.

34. Babak Ehteshami Bejnordi *et al.*, « Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer », *JAMA*, 2017, 318, p. 2199.

35. Varun Gulshan *et al.*, « Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs », *JAMA*, 2016, 316, p. 2402.

36. Mary Beth Massat, « A promising future for AI in breast cancer screening », *Applied Radiology*, 2018, 47, p. 22 ; Alejandro Rodriguez-Ruiz *et al.*, « Stand-alone artificial intelligence for breast cancer detection in mammography : Comparison with 101 radiologists », *Journal of the National Cancer Institute*, 2019, 111, p. 916.

37. Apgar Score, Medline Plus, https://medlineplus.gov/ency/article/003402.htm, consulté le 4 février 2020.

38. L. R. Foster *et al.*, « The interrater reliability of Apgar Scores at 1 and 5 minutes », *Journal of Investigative Medicine*, 2006, 54 (1), p. 293.

39. Warren J. McIsaac *et al.*, « Empirical validation of guidelines for the management of pharyngitis in children and adults », *JAMA*, 2004, 291, p. 1587.

40. Emilie A. Ooms *et al.*, « Mammography : Interobserver variability in breast density assessment », *Breast*, 2007, 16, p. 568.

41. Frances P. O'Malley *et al.*, « Interobserver reproducibility in the diagnosis of flat epithelial atypia of the breast », *Modern Pathology*, 12006, 9, p. 172.

42. Voir Ahmed Aboraya *et al*., « The reliability of psychiatric diagnosis revisited », *Psychiatry (Edgmont)*, 2006, 3, p. 41. Pour une synthèse, voir N. Kreitman, « The reliability of psychiatric diagnosis », *Journal of Mental Science*, 1961, 107, p. 876-886.

43. Ahmed Aboraya *et al*., « The reliability of psychiatric diagnosis revisited », art. cit., p. 43.

44. C. H. Ward *et al*., « The psychiatric nomenclature : Reasons for diagnostic disagreement », *Archives of General Psychiatry*, 1962, 7, p. 198.

45. Ahmed Aboraya *et al*., « The reliability of psychiatric diagnosis revisited », art. cit.

46. Samuel M. Lieblich, David J. Castle, Christos Pantelis, Malcolm Hopwood, Allan Hunter Young, Ian P. Everall, « High heterogeneity and low reliability in the diagnosis of major depression will impair the development of new drugs », *British Journal of Psychiatry Open*, 2015, 1 (2), p. e5-e7.

47. *Ibid*.

48. Voir Elie Cheniaux *et al*., « The diagnoses of schizophrenia, schizoaffective disorder, bipolar disorder and unipolar depression : Interrater reliability and congruence between DSM-IV and ICD-10 », *Psychopathology*, 2009, 42, p. 296-298, en particulier p. 293 ; et Michael Chmielewski *et al*., « Method matters : Understanding diagnostic reliability in DSM-IV and DSM-5 », *Journal of Abnormal Psychology*, 2015, 124, p. 764.

49. Ahmed Aboraya *et al*., « The reliability of psychiatric diagnosis revisited », art. cit., p. 47.

50. *Ibid*.

51. Voir Michael Chmielewski *et al*., « Method matters : Understanding diagnostic reliability in DSM-IV and DSM-5 », art. cit.

52. Voir par exemple Helena Chmura Kraemer *et al*., « DSM-5 : How reliable is reliable enough ? », *American Journal of Psychiatry*, 2012, 169, p. 13-15.

53. S. M. Lieblich *et al*., « High heterogeneity and low reliability in the diagnosis of major depression will impair the development of new drugs », art. cit.

54. *Ibid.*, p. e5.

55. *Ibid*.

56. *Ibid.*, p. e6.

57. Ahmed Aboraya *et al*., « The reliability of psychiatric diagnosis revisited », art. cit., p. 47.

58. *Ibid*.

59. *Ibid*.

60. On trouvera des remarques prudentes dans Christopher Worsham, Anupam B. Jena, « The art of evidence-based medicine », *Harvard Business Review*, 30 janvier 2019.

# CHAPITRE 23

## Évaluation de la performance : le choix de l'échelle

1. Jena McGregor, « Study finds that basically every single person hates performance reviews », *Washington Post*, 27 janvier 2014.

2. La transformation numérique que connaissent de nombreuses organisations pourrait créer ici des possibilités nouvelles. En théorie, les entreprises peuvent désormais collecter de grandes quantités d'informations granulaires en temps réel sur la performance de chaque travailleur. Ces données rendent possibles des évaluations de la performance entièrement algorithmiques. Nous nous intéresserons cependant ici aux postes pour lesquels on ne peut pas éliminer totalement le jugement de la mesure de la performance. Voir E. D. Pulakos, R. Mueller-Hanson, S. Arad, « The evolution of performance management : Searching for value », *Annual Review of Organizational Psychology and Organizational Behavior*, 2018, 6, p. 249-271.

3. S. E. Scullen, M. K. Mount, M. Goff, « Understanding the latent structure of job performance ratings », *Journal of Applied Psychology*, 2000, 85, p. 956-970.

4. Une petite partie – 10 % de la variance dans certaines études – est ce que les chercheurs appellent la *perspective de l'évaluateur* ou l'*effet de niveau*, au sens de niveau dans l'organisation (pas de bruit de niveau). La perspective de l'évaluateur témoigne du fait que dans l'évaluation de la même personne, le supérieur hiérarchique diffère systématiquement d'un pair, et un pair d'un subordonné. Si l'on fait une interprétation charitable des résultats de systèmes d'évaluation à 360 degrés, on pourrait dire que ce n'est pas du bruit. Si des gens placés à différents niveaux de l'organisation voient systématiquement des facettes différentes de la performance de la même personne, il est légitime que leur jugement sur cette personne diffère systématiquement et que leurs évaluations le reflètent.

5. S. E. Scullen, M. K. Mount, M. Goff, « Understanding the latent structure of job performance ratings », art. cit. ; C. Viswesvaran, D. S. Ones, F. L. Schmidt, « Comparative analysis of the reliability of job performance ratings », *Journal of Applied Psychology*, 1996, 81, p. 557-574 ; G. J. Greguras, C. Robie, « A new look at within-source interrater reliability of 360-degree feed-back ratings », *Journal of Applied Psychology*, 1998, 83, p. 960-968 ; G. J. Greguras, C. Robie, D. J. Schleicher, M. A. Goff, « A field study of the effects of rating purpose on the quality of multisource ratings », *Personnel Psychology*, 2003, 56, p. 1-21 ; C. Viswesvaran, F. L. Schmidt, D. S. Ones, « Is there a general factor in ratings of job performance ? A meta-analytic framework for disentangling substantive and error influences », *Journal of Applied Psychology*, 2005, 90, p. 108-131 ; et B. Hoffman, C. E. Lance, B. Bynum, W. A. Gentry, « Rater source effects are alive and well after all », *Personnel Psychology*, 2010, 63, p. 119-151.

6. K. R. Murphy, « Explaining the weak relationship between job performance and ratings of job performance », *Industrial and Organizational Psychology*, 2008, 1, p. 148-160, en particulier p. 151.

7. Dans la discussion sur les sources du bruit, nous ne mentionnons pas la possibilité que les biais systématiques dans l'évaluation de certains employés ou certaines catégories d'employés produisent une erreur partagée. Aucune des études que nous avons pu lire sur la variabilité des évaluations de performance ne les comparait avec une « vraie » performance évaluée de l'extérieur.

8. E. D. Pulakos, R. S. O'Leary, « Why is performance management broken ? », *Industrial and Organizational Psychology*, 2011, 4, p. 146-164 ; M. M. Harris, « Rater motivation in the performance appraisal context : A theoretical framework », *Journal of Management*, 1994, 20, p. 737-756 ; et K. R. Murphy, J. N. Cleveland, *Understanding Performance Appraisal : Social, Organizational, and Goal-Based Perspectives*, Sage, 1995.

9. G. J. Greguras *et al*., « A field study of the effects of rating purpose on the quality of multisource ratings », art. cit.

10. P. W. Atkins, R. E. Wood, « Self-versus others' ratings as predictors of assessment center ratings : Validation evidence for 360-degree feed-back programs », *Personnel Psychology*, 2002, 55 (4), p. 871-904.

11. *Ibid.*

12. Olson et Davis, cité par Peter G. Dominick, « Forced ranking : Pros, cons and practices », *in* James W. Smither, Manuel London (dir.), *Performance Management : Putting Research into Action*, Jossey-Bass, 2009, p. 411-443.

13. Peter G. Dominick, « Forced ranking : Pros, cons and practices », art. cit.

14. La figure est adaptée de Richard D. Goffin, James M. Olson, « Is it all relative ? Comparative judgments and the possible improvement of self-ratings and ratings of others », *Perspectives on Psychological Science*, 2011, 6 (1), p. 48-60.

15. Barry R. Nathan, Ralph A. Alexander, « A comparison of criteria for test validation : A meta-analytic investigation », *Personnel Psychology*, 1988, 41 (3), p. 517-535.

16. M. Buckingham, A. Goodall, « Reinventing performance management », *Harvard Business Review*, 1er avril 2015, p. 1-16.

17. Corporate Leadership Council, cité *in* S. Adler *et al*., « Getting rid of performance ratings : Genius or folly ? A debate », *Industrial and Organizational Psychology*, 2016, 9, p. 219-252.

18. E. D. Pulakos, R. Mueller-Hanson, S. Arad, « The evolution of performance management : Searching for value », art. cit., p. 250.

19. A. Tavis, P. Cappelli, « The performance management revolution », *Harvard Business Review*, octobre 2016, p. 1-17.

20. Frank J. Landy, James L. Farr, « Performance Rating », *Psychological Bulletin*, 1980, 87 (1), p. 72-107.

21. D. J. Woehr, A. I. Huffcutt, « Rater training for performance appraisal : A quantitative review », *Journal of Occupational and Organizational Psychology*, 1994, 67, p. 189-205 ; S. G. Roch, D. J. Woehr, V. Mishra, U. Kieszczynska, « Rater training revisited : An updated meta-analytic review of frame-of-reference training », *Journal of Occupational and Organizational Psychology*, 2012, 85, p. 370-395 ; et M. H. Tsai, S. Wee, B. Koh, « Restructured frame-of-reference training improves rating accuracy », *Journal of Organizational Behavior*, 2019, 40 (6), p. 1-18.

22. Le panel de gauche est adapté de Richard Goffin et James M. Olson, « Is it all relative ? Comparative judgments and the possible improvement of self-ratings and ratings of others », *Perspectives on Psychological Science*, 2011, 6 (1), p. 48-60.

23. S. G. Roch *et al.*, « Rater training revisited : An updated meta-analytic review of frame-of-reference training », art. cit.

24. Ernest O'Boyle, Herman Aguinis, « The best and the rest : Revisiting the norm of normality of individual performance », *Personnel Psychology*, 2012, 65 (1), p. 79-119 ; et Herman Aguinis, Ernest O'Boyle, « Star performers in Twenty-First Century organizations », *Personnel Psychology*, 2014, 67 (2), p. 313-350.

# CHAPITRE 24

## RECRUTEMENT : LA STRUCTURATION DES JUGEMENTS COMPLEXES

1. A. I. Huffcutt, S. S. Culbertson, « Interviews », *in* S. Zedeck (dir.), *APA Handbook of Industrial and Organizational Psychology*, American Psychological Association, 2010, p. 185-203.

2. N. R. Kuncel, D. M. Klieger, D. S. Ones, « In hiring, algorithms beat instinct », *Harvard Business Review*, 2014, 92 (5), p. 32.

3. R. E. Ployhart, N. Schmitt, N. T. Tippins, « Solving the supreme problem : 100 years of selection and recruitment at the *Journal of Applied Psychology* », *Journal of Applied Psychology*, 2017, 102, p. 291-304.

4. M. McDaniel, D. Whetzel, F. L. Schmidt, S. Maurer, « Meta-analysis of the validity of employment interviews », *Journal of Applied Psychology*, 1994, 79, p. 599-616 ; A. Huffcutt, W. Arthur, « Hunter and hunter (1984) revisited : Interview validity for entry-level jobs », *Journal of Applied Psychology*, 1994, 79, p. 2 ; F. L. Schmidt, J. E. Hunter, « The validity and utility of selection methods in personnel psychology : Practical and theoretical implications of 85 years of research findings », *Psychology Bulletin*, 1998, 124 p. 262-274 ; et F. L. Schmidt, R. D. Zimmerman, « A counterintuitive hypothesis about employment interview validity and some supporting evidence », *Journal of Applied Psychology*, 2004, 89, p. 553-561. Notez que les validités sont plus élevées quand certains sous-ensembles d'études sont pris en considération, notamment quand l'étude utilise des évaluations de performance créées spécialement, plutôt que les évaluations existantes de l'organisation.

5. S. Highhouse, « Stubborn reliance on intuition and subjectivity in employee selection », *Industrial and Organizational Psychology*, 2008, 1, p. 333-342 ; D. A. Moore, « How to improve the accuracy and reduce the cost of personnel selection », *California Management Review*, 2017, 60, p. 8-17.

6. L. A. Rivera, « Hiring as cultural matching : The case of elite professional service firms », *American Sociology Review*, 2012, 77, p. 999-1022.

7. F. L. Schmidt, R. D. Zimmerman, « A counterintuitive hypothesis about employment interview validity and some supporting evidence », art. cit. ; Timothy A. Judge, Chad A. Higgins, Daniel M. Cable, « The employment interview : A review of recent research and

recommendations for future research », *Human Resource Management Review*, 2000, 10, p. 383-406 ; et A. I. Huffcutt, S. S. Culbertson, W. S. Weyhrauch, « Employment interview reliability : New meta-analytic estimates by structure and format », *International Journal of Selection and Assessment*, 2013, 21, p. 264-276.

8. M. R. Barrick *et al.*, « Candidate characteristics driving initial impressions during rapport building : Implications for employment interview validity », *Journal of Occupational and Organizational Psychology*, 2012, 85, p. 330-352 ; M. R. Barrick, B. W. Swider, G. L. Stewart, « Initial evaluations in the interview : Relationships with subsequent interviewer evaluations and employment offers », *Journal of Applied Psychology*, 2010, 95, p. 1163.

9. G. L. Stewart, S. L. Dustin, M. R. Barrick, T. C. Darnold, « Exploring the handshake in employment interviews », *Journal of Applied Psychology*, 2008, 93, p. 1139-1146.

10. T. W. Dougherty, D. B. Turban, J. C. Callender, « Confirming first impressions in the employment interview : A field study of interviewer behavior », *Journal of Applied Psychology*, 1994, 79, p. 659-665.

11. J. Dana, R. Dawes, N. Peterson, « Belief in the unstructured interview : The persistence of an illusion », *Judgment and Decision Making*, 2013, 8, p. 512-520.

12. Nathan R. Kuncel *et al.*, « Mechanical versus clinical data combination in selection and admissions decisions : A meta-analysis », *Journal of Applied Psychology*, 2013, 98 (6), p. 1060-1072.

13. Laszlo Bock, entretien avec Adam Bryant, *The New York Times*, 19 juin 2013. Voir aussi Laszlo Bock, *Work Rules ! Insights from Inside Google That Will Transform How You Live and Lead*, Hachette, 2015.

14. C. Fernandez-Araoz, « Hiring without firing », *Harvard Business Review*, 1er juillet 1999.

15. Pour un guide accessible de l'entretien structuré, voir Michael A. Campion, David K. Palmer, James E. Campion, « Structuring employment interviews to improve reliability, validity and users' reactions », *Current Directions in Psychological Science*, 1998, 7 (3), p. 77-82.

16. J. Levashina, C. J. Hartwell, F. P. Morgeson, M. A. Campion, « The structured employment interview : Narrative and quantitative review of the research literature », *Personnel Psychology*, 2014, 67, p. 241-293.

17. M. McDaniel *et al.*, « Meta-analysis of the validity of employment interviews », art. cit. ; A. Huffcutt, W. Arthur, « Hunter and hunter (1984) revisited : Interview validity for entry-level jobs », art. cit. ; F. L. Schmidt, J. E. Hunter, « The validity and utility of selection methods in personnel psychology : Practical and theoretical implications of 85 years of research findings », art. cit. ; F. L. Schmidt, R. D. Zimmerman, « A counterintuitive hypothesis about employment interview validity and some supporting evidence », art. cit.

18. F. L. Schmidt, J. E. Hunter, « The validity and utility of selection methods in personnel psychology : Practical and theoretical implications of 85 years of research findings », art. cit.

19. D. Kahneman, *Thinking, Fast and Slow*, *op. cit.*, p. 229 ; trad. fr. *Système 1/Système 2, Les deux vitesses de la pensée*, *op. cit.*, chapitre 21.

20. N. R. Kuncel, D. M. Klieger, D. S. Ones, « In hiring, algorithms beat instinct », art. cit. Voir aussi Michael A. Campion, David K. Palmer, James E. Campion, « Structuring employment interviews to improve reliability, validity and users' reactions », art. cit.

21. J. Dana, R. Dawes, N. Peterson, « Belief in the unstructured interview : The persistence of an illusion », art. cit.

## CHAPITRE 25
### Le protocole d'évaluations intermédiaires

1. Daniel Kahneman, Dan Lovallo, Olivier Sibony, « A structured approach to strategic decisions : Reducing errors in judgment requires a disciplined process », *MIT Sloan Management Review*, 2019, 60, p. 67-73.

2. Andrew H. Van De Ven, Andre Delbecq, « The effectiveness of nominal, Delphi, and interacting group decision making processes », *Academy of Management Journal*, 1974, 17 (4), p. 605-621. Voir aussi le chapitre 21.

# PARTIE VI
# Le bruit dans les jugements prédictifs

1. Kate Stith, Jose A. Cabranes, *Fear of Judging : Sentencing Guidelines in the Federal Courts*, University of Chicago Press, 1998, p. 177.

## CHAPITRE 26
### Les coûts de la réduction du bruit

1. Albert O. Hirschman, *The Rhetoric of Reaction : Perversity, Futility, Jeopardy*, Belknap Press, 1991.

2. Kate Stith, Jose A. Cabranes, *Fear of Judging*, *op. cit.*

3. Voir par exemple *Three Strikes Basics*, Stanford Law School, https://law.stanford.edu/stanford-justice-advocacy-project/three-strikes-basics/.

4. 428 U.S. 280, 1976.

5. Cathy O'Neil, *Weapons of Math Destruction : How Big Data Increases Inequality and Threatens Democracy*, Crown Publishers, 2016.

6. Will Knight, « Biased algorithms are everywhere, and no one seems to care », *MIT Technology Review*, 12 juillet 2017.

7. Jeff Larson, Surya Mattu, Lauren Kirchner, Julia Angwin, « How we analyzed the COMPAS recidivism algorithm », *ProPublica*, 23 mai 2016, www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. La présence de biais dans cet exemple est contestée, et des définitions différentes du biais peuvent conduire à des conclusions opposées. Sur ce cas et plus largement sur la définition du biais algorithmique, voir la note 10 ci-dessous.

8. Aaron Shapiro, « Reform predictive policing », *Nature*, 2017, 541 (7638), p. 458-460.

9. Si cette préoccupation refait surface dans le contexte des modèles basés sur l'IA, elle n'est pas propre à l'IA. Dès 1972, Paul Slovic notait que la modélisation de l'intuition préserverait et renforcerait, et peut-être même amplifierait, les biais cognitifs existants. Paul Slovic, « Psychological study of human judgment : Implications for investment decision making », *Journal of Finance*, 1972, 27, p. 779. Voir également la note 10 ci-dessous.

10. Pour une introduction à ce débat dans le contexte de la controverse sur l'algorithme de prédiction de la récidive Compas, voir Larson *et al.*, « How we analyzed the COMPAS recidivism algorithm », art. cit. ; William Dieterich *et al.*, « COMPAS risk scales : Demonstrating accuracy equity and predictive parity », Northpointe Inc., 8 juillet 2016, http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf ; Julia Dressel, Hany Farid, « The accuracy, fairness, and limits of predicting recidivism », *Science Advances*, 2018, 4 (1), p. 1-6 ; Sam Corbett-Davies *et al.*, « A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear », *The Washington Post*, 17 octobre 2016 ; Alexandra Chouldechova, « Fair prediction with disparate impact : A study of bias in recidivism prediction instruments », *Big Data*, 2017, 153, p. 5 ; et Jon Kleinberg, Sendhil Mullainathan, Manish Raghavan, « Inherent trade-offs in the fair determination of risk scores », Leibniz International Proceedings in Informatics, janvier 2017.

# CHAPITRE 27
## UNE AFFAIRE DE DIGNITÉ

1. Voir par exemple la Convention européenne des droits de l'homme, article 6.

2. Tom R. Tyler, *Why People Obey the Law*, Princeton University Press, 2020, 2e éd.

3. *Cleveland Bd. of Educ. v. LaFleur*, 414 U.S. 632, 1974.

4. Laurence H. Tribe, « Structural due process », *Harvard Civil Rights–Civil Liberties Law Review*, 1975, 10 (2), p. 269.

5. Kate Stith, Jose A. Cabranes, *Fear of Judging : Sentencing Guidelines in the Federal Courts*, University of Chicago Press, 1998, p. 177.

6. Voir par exemple Philip K. Howard, *The Death of Common Sense : How Law Is Suffocating America*, Random House, 1995 ; et Philip K. Howard, *Try Common Sense : Replacing the*

*Failed Ideologies of Right and Left*, W. W. Norton and Company, 2019.

# CHAPITRE 28
## DES RÈGLES OU DES PRINCIPES ?

1. « Hate speech », Facebook : Community Standards, www.facebook.com/communitystandards/hate_speech.
2. Andrew Marantz, « Why Facebook can't fix itself », *The New Yorker*, 12 octobre 2020.
3. Jerry L. Mashaw, *Bureaucratic Justice*, Yale University Press, 1983.
4. David M. Trubek, « Max Weber on law and the rise of capitalism », *Wisconsin Law Review*, 1972, 720, p. 733, note 22 (citant Max Weber, *The Religion of China : Confucianism and Taoism*, The Free Press, 1951, p. 149).

# Index


Aboraya, Ahmed 296

accidents d'échantillonnage 131

accord sur des principes que sur des règles 362

ACE-V 259, 262

action 32

actualisation continue 276

actuarielle 117

adaptation 339

*Adult Decision Making Competence scale* 243

affirmations absurdes 92

agences de renseignement 278

agrégation 157, 271

- de jugements indépendants 233, 263, 303, 319, 384

aides aux personnes handicapées 365

aisance 35

algorithmes 128, 231, 292, 346, 382

- de « police prédictive » 347
- de recrutement 138
- transparents 348

Amazon Mechanical Turk 222

ambiguïté 208, 330

analyse

coût-bénéfice 386

de l'ADN 259

angine bactérienne 287

angiographie coronarienne 289

anniversaire du prévenu 22

anticipation de désaccord limité 47, 372

Apgar (score) 293-295, 346, 362, 383

Apgar, Virginia 293

apprentissage automatique 128

approche causale 40

aptitudes mentales générales (*general mental ability* ou GMA) 239

architecture du choix (*nudges*) 248

Aristote 237

Armstrong, J. Scott 272

ascenseur, charge maximale 70

asile

demande d' 94

politique 12, 22

assurances 27-30, 32

astéroïde 65

astronomes 65

attentat(s)

de Madrid 257

du World Trade Center 160

attention sélective 49

audit de bruit 27, 29, 60, 73, 195, 231, 374, 393

des juges fédéraux 73
Austin, Williams 21
automaticité 24
autopsies 36
autoréalisation 101
aventurier 183
aversion
à la perte 251
aux algorithmes 140
avocats 235
Baron, Jonathan 244
batterie de tests prédictifs 284
Bentham, Jeremy 92, 237
Bertillon, Alphonse 258
bertillonnage 258
besoin
de cognition (*need for cognition*, ou NFC) 243
d'explication 226
bêta perpétuel 278
biais 9
algorithmique 138-139, 346
charisme explicatif 379
correction 248
d'ancrage 176, 202, 249
dans le recrutement 314
de conclusion 174
de confirmation 175, 249, 261-263
de confirmation (criminalistique) 261-262
de désirabilité 175

du planificateur 168, 248, 251

du *statu quo* 251

formation à surmonter 249

partagé 179

point aveugle 252, 267

politiques 170

psychologiques (effets) 167, 179, 282, 378

psychologiques (étude) 168

rétrospectif 226

statistique 167, 282

Big Five (modèle) 121, 215

Bill, le comptable jazzman 170-172, 182-183

BIN (*bias, information, noise*) (modèle) 281

BI-RADS (*Breast Imaging Reporting and Data System*) 295

BMW repeinte 197

Bock, Laszlo 319, 321

bonne humeur 91

*boosting* 249

braquage 20

de banque 21, 24

*Breast Imaging Reporting and Data System* (BI-RADS) 295

brevets 13

Breyer, Stephen 23

Brier, Glenn (score de) 276-277, 280

Brier, Glenn W. 276

bruit 10

absence de 139

de niveau 77, 218, 299, 301, 306, 344, 376

de pattern 78-79, 211, 218, 220, 301, 315, 376

existence 267
hors la loi 368
inattention au 34, 228, 379
occasionnel 81, 90, 218, 264, 269, 272, 294, 301, 315, 377
optimal 337
rapport coûts-bénéfices (réduction) 342
réduction du 26, 67
systémique 26, 29, 373
*bullshit*, réceptivité 92
Cabranes, José 24, 338
cadre(s)
de référence commun 309
dirigeantes (évaluation) 115
candidat
à l'embauche 55
apparence physique 180, 315
carabine 9
caractère consciencieux 239
cas
associations 212
nombreux 73
unique 73
cas-ancres 309
cascade
de biais 263
de pression sociale 106
informationnelle 104
casier judiciaire 21
causalité 158

Centor (score) 294

Centor, Robert 295

Cérès 65

chaîne causale 158

changement

climatique 173

par le protocole d'évaluation 334

check-list 252

« chiffre magique de sept, Le » (Miller) 189-190

choix stratégiques 55

Choses Normales, vallée des 154, 160-162, 164, 226

chronomètre 44

cibles de dos 10, 57

circonstances particulières 24

classe de référence 328, 382

clémence 352

cliché du leader 245

coefficient 117

kappa 288

« optimaux » 130

coefficient de corrélation 112

et pourcentage de concordance (PC) 113

cohérence 31, 57

dans l'arbitraire 203

globale 209

Comité des avocats pour les droits humains (Human Rights First) 19

commentaires en ligne 102

Commission des peines (États-Unis) 21, 23, 26, 75

*community standards* 364-365

Compas (algorithme) 347
compétence 235
comportements opportunistes 339
compréhension et prédiction 158
concordance interjuges 288
concours de vins 48
concurrence 32
confiance 238
    que le « principal » accorde à son « agent » 367
conflit 35
    d'intérêts 170
consensus 35
« consensus » des analystes financiers 273
constance
    irréfléchie 126, 378
    sans consensus 224
consultants 27
contexte 185
contournement(s) 366
contrefactuel 41, 86
    contrefaçon (chèques) 20
corrélation(s) 313
    contre-validée 130
    en sciences sociales 157
    et causalité 158
Corrigan, Bernard 129-131
Cour suprême 23, 25, 128, 272, 344-345, 351, 353
coût(s)
    de l'erreur 70, 367, 373

de la décision 367
de l'élimination du bruit 304, 310, 338
des processus d'évaluation 308
du bruit 30, 70, 375
et faisabilité des efforts de réduction du bruit 342
symétriques 70
covariance 111
Covid-19 41, 287
Cowgill, Bo 138
créativité 239
et motivation 358
crédulité 92
*Criminal Sentences : Law Without Order* 20
critères d'évaluation (définitions) 327
critiques
de cinéma 31
gastronomiques 55
Curry, Stephen 84
dactylotechnie et dactyloscopie 259
Dartmouth Atlas Project 287
Darwin, Charles 88
Dawes, Robyn 129-132
décision(s)
bureaucratique 369
non conclusive 259
récurrentes 38
répétée 368
singulières 38, 385
Déclaration universelle des droits de l'homme 363, 369-370

décomposition 319
degré d'obstruction 289
Deloitte (cabinet de conseil) 308
Delphi (méthode) 274
démotivation 340
déni
    de l'ignorance 150
    du biais 267
déontologique 92
dépression 295
désaccord 35, 285
descriptions de poste 320
détournement de fonds 20, 24
*Deux siècles de rhétorique réactionnaire* (Hirschman) 343
développeurs de logiciels 12
diabète 287
*Diagnostic and Statistical Manual of Mental Disorders*, 3e édition (*DSM-III*) 296
*Diagnostic and Statistical Manual of Mental Disorders*, 4e édition (*DSM-IV*) 296
*Diagnostic and Statistical Manual of Mental Disorders*, 5e édition (*DSM-5*) 296
diagnostic(s) 286
différence(s)
    dans les niveaux de compétence 212
    entre estimations 30
    individuelles dans les biais des juges 180
    interindividuelles sur le bruit de niveau 225
dignité 339
directeurs financiers (étude) 271
directive(s) 23, 233, 286-287, 320, 338
    indicatives 25

obligatoire 25
discrimination(s) 137
intentionnelle 348
non intentionnelle 348
par les algorithmes 138
disparité(s) 23
de compétences 288
dispersion 10
dissuasion 339
distribution
de cocaïne et d'héroïne 24
des performances réelles 307
distribution normale 307
gaussienne courbe en cloche 60
diversité 283-285, 375
divulgation séquentielle linéaire 268
dommages 195
dommages et intérêts 199
punitifs 196
données objectives (évaluer la performance) 300
doute raisonnable 43
Dreyfus, Alfred 258
Dror, Itiel 260-262, 264, 266-269
Dylan, Bob 184, 191
Ebola (virus) 38
écart-type 45, 60
échelle(s) 194, 299
ambiguës 196, 206
choix des 165

comparatives 192
comportementale ancrée 309, 321
de cas 309, 385
de rapport 202
de réponse 195
d'intensité 184
nombre de catégories 189
partagée calibrée par une vision externe 233
qualitatives 184
effets
de cascade 104
de halo 178, 304, 306, 329
de séquence 94
pervers 343
électeurs démocrates ou républicains 102
élément irrecevable 179
emploi 39
empreintes digitales 258
emprisonnement 21
endométriose 290
*énergie nucléaire* 174
ensembles 163
entretien
comportemental structuré 321
structuré 297
entretien d'embauche 312
validité prédictive 313
environnements concurrentiels 31
épidémie 55

équipe
de football, défaite 22
de professionnels 284
travail en 280
équité de traitement 56
équivalences 182
création d’ 165, 182
d’intensité 184
erreur(s) 9, 31
aléatoires 33
asymétrie 266
composantes 218
d’attribution 226
de niveau 77
de pattern 79
de pondération 174
du parieur 94
équation 66
globale 59
judiciaires 264
jugement (explication) 226
quadratique moyenne 63
esprit, ouverture 235, 244, 279, 380
estimation
de GPA 186
discussion 331
État de droit 20-21, 351
étude(s)
de cas 40

d’impact 253
Eugene (Oregon) 121, 129
« Évaluation de la pensée critique de Halpern » (*Halpern Critical Thinking Assessment*) 243
évaluation(s) 302
à 360 degrés 301
calculs stratégiques 302
de la performance 299
des professeurs 176
du jugement 53
intermédiaires 320
« sans chiffres » 309
événements
attendus 160
s’expliquent tout seuls 161
éventail des scores de GMA 240
exactitude 70
excès
de cohérence 177, 179, 316-317, 330, 383
de confiance 146, 249-251
de précision 271
d’optimisme 271
exclusion 259
expérience 36, 238
*Expert Political Judgment* (Tetlock) 146
experts 35, 146, 236
d’estime 235-236, 380
en sinistres 28
expression individuelle 381
expulsion 155

extorsion de fonds 21

Facebook 343, 364

facteurs occasionnels non pertinents 87

faim 22

fatigue 84, 93, 291

Faulds, Henry 258

faux négatifs 266

faux positifs 265-266

Federal Bureau of Investigation (FBI) 257, 259, 263-268

feed-back à 360 degrés, inconvénients 303

fiabilité intraobservateur 264

financements à des projets de recherche 55

Fingerprint Society 267

foire agricole 88

*forced ranking* (classement forcé) 304

Forgas, Joseph 90-91

formation 237, 280

- au cadre de référence 309

formules 128

foule

- intérieure 87-88
- sagesse 88, 99, 103

*Fragile Families and Child Wellbeing Study* 155, 157-158

Frankel, Marvin 19-21, 23-26, 55-56, 75, 140, 337, 343

Frankfurt, Harry 92

Freud, Sigmund 120

gain

- obtenu 88
- résultant de la réduction du bruit 67

Galton, sir Francis 88, 258

Gambardi (cas) 50-53, 172-173, 183, 278, 404

Gambardi, Michael 48-52, 169, 172-173, 183

Gates, Bill 241

Gauss, Carl Friedrich 63-65

General Electric 304

gènes (membres d'une même fratrie) 112

geste, répétition 44

Glover, Joan 194-198, 202-203

Goldberg, Lewis 121-123, 125-126

Good Judgment Project 274, 285

GoodSell (exemple de réduction de bruit) 60-62, 67, 69-70

Google 319-322

gourous du management 40

goût 31

*grit*, « niaque » 156, 239

groupes 99

- délibération 285
- polarisation 107

*Guerre et paix* 185

*Halpern Critical Thinking Assessment* 243

Haran, Uriel 244

Harvard 25, 202

Harvard Business School 48

Havel, Václav 343

HEC Paris 250

Hertwig, Ralph 89

Herzog, Stefan 89

heuristique(s) 167

de disponibilité 173, 250
de la confiance 238
de l'affect 175
et biais 167
Hirschman, Albert 343
historiens 40
Hitchcock, Alfred 270
Hoffman, Paul 121
hors échantillon 130
Howard, Philip 359
*Human Resources Management* 303
Human Rights First (Comité des avocats pour les droits humains) 20
humeur 90
négative 91
hygiène de la décision 232, 247, 254, 335, 381
gains attendus 233
hypothèse de l'indignation 197
identification 259
ignorance objective 143, 314, 377
illusion
d'accord 33-34, 380
de validité 119, 186
*implementation standards* 364-365
impôts 356
impressions, premières 91, 316
inadaptation de la nomenclature 296
incertitude irréductible 145
incohérence 208
incompétence 231

indépendance 319
de jugement 103, 268
indésirable *vs* souhaitable 31
indications caloriques 178
indices contradictoires 208
indignation 195, 197, 199
individualité 225
inflation des notes 304
inflexibilité 355
influence sociale 100
information(s)
imparfaite 145
parasites 268
pour édicter des règles 363
séquencement 232, 268-269
supplémentaires sur l'affaire 261
injustice du biais et du bruit 369-370
Innocence Project 264
innovation 31
insensibilité à l'étendue 169
intégration informelle 49
intelligence 235, 239, 278-279
artificielle (IA) 117, 128
collective 326
cristallisée 239
fluide 240
intention de sanctionner 195, 199
interactions 225
médicamenteuses 289

statisticien 79
interprétations alternatives 209
suppression des 209
intervalle de confiance 272
intuition(s) 143, 383
dans la prise de décision managériale 143
sur l'erreur 68-69
invérifiable 51
jambes cassées 134-135, 332
jazz 171
jeux 44
sérieux 249
*Joan Glover vs General Assistance* (exemple) 194-198, 200-203
Jobs, Steve 241
*Journal of Applied Psychology* 313
jugement(s)
absolus 305
clinique 116
comparatif 191, 385
complémentaires 283
définitif 35
définition 43, 371
de fréquence 173
de probabilité 51
différé 319, 322, 384
documentation 268
humain 10
moraux 92
normaux 226

peur 338

préalable 175

relatifs 305, 309, 328

unidimensionnels 379

jugements évaluatifs 55, 372

dans les décisions 71

jugements prédictifs 45, 111, 372

dans les décisions 71

juges

définition 44, 371

meilleurs 235

pour enfants 22

Julie (exemple) 185-189, 207-209, 212, 214, 403-406

jury(s) 205

délibérant 108

*nullification* (invalidation par le jury) 366

statistiques 107

justesse 44

justice

bureaucratique 366

de cadi 368

Kahana, Michael 96-97

Kahneman, Daniel 195, 322, 324

Kant, Emmanuel 93, 237

Kasdan, Lawrence 174-175

Kennedy, Edward M. 23

Kennedy, John F. 23

Keynes, John Maynard 276

Kuncel, Nathan 36, 125-126

LaFleur, Jo Carol 353-354

lancer franc 84

lavage des mains 254, 291

lésion(s)

    cutanées 290

    mammaire 287, 289

Lewis, Michael 140

liberté sous caution 12, 135, 224

Lieblich, Samuel 296

« loi des trois prises » ou « de la troisième infraction » (*Three strikes and you're out*) 344, 366

loterie 28-29

Lucas, George 174-175

*machine learning* 128

Macy, Michael 102

maison (achat) 39

mammographies 290

Mapco/Roadco (exemple) 324-325, 327-329, 331, 333-334

    acquisition stratégique 324

*Marchand de Venise, Le* (Shakespeare) 352

marché 32

    de prédictions 273

mariage 39

marijuana 21

Mashaw, Jerry 366, 369

matrice de handicap 365

Mayfield, Brandon 257, 263, 265-267

McLanahan, Sara 154, 158

*mean squared error* (MSE) 63-72, 156, 218, 373-374

mécanique 117

médecine 11, 286
médiane 63
Meehl, Paul 118-121, 123, 134, 140, 151
mélanome 43, 290
Mellers, Barbara 148, 244, 274, 280-281
mémoire 96
  sélective 49
mesure 43
  instrument 43
météorologie 94, 249
méthode(s)
  d'agrégation 272
  des carrés 63
  économétriques 87
milliardaires 241
mini-Delphi, estimation-discussion-estimation 274
*Minority Report* 149
*minuties* 262
modèle de personnalité en cinq facteurs (Big Five) 121, 215
modèle(s)
  « comme si » 122
  du juge 121-122
  équipondéré 129
  frugaux 132, 378
modèles linéaires
  aléatoires 126
  impropres 129
modérateurs 364-365
mondes parallèles 100

*Moneyball* (Lewis) 140
Moore, Don 274
Morewedge, Carey 249
moyenne
    arithmétique 64
    de *n* jugements 272
    des valeurs absolues des différences 76
    d'un groupe sélectionné 273
Muchnik, Lev 103
Mullainathan, Sendhil 135-138, 149
Nash, Steve 84
National Basketball Association (NBA) 84
négociation 91, 176
néphropathie 289
niveau
    de désaccord 47
    professionnel 239
normes et pratiques 237
note masquée 36
O'Neal, Shaquille 84
O'Neil, Cathy 346
Obama, Barack 38-41
Obermeyer, Ziad 149
objections 26
    à la prédiction mécanique 140
observateurs de la décision (choix) 232, 252, 399
octroi de brevets 221
œnologues 31
OMB A4 (circulaire) 253

*On Bullshit* (Frankfurt) 92, N17
options 55
organisations 32
    flexibles 303
originalité 125
oukases personnels 20
Pashler, Harold 87-90
pathologie 288
patinage artistique 55
pattern(s) 207
    personnel 213
pause déjeuner 22
PCAST (President's Council of Advisors on Science and Technology) 265, 267-268
P-DG 48, 241
peine
    criminelle 19
    de mort obligatoire 344
    individualisation 19, 345
    juste 38
    pénale, finalité 78
Pennycook, Gordon 92
pensée
    causale 159, 380
        inaudibilité du bruit 228
    probabiliste 42
    statistique 159, 380, 382
performance
    absolue et performance relative 306
    mnésique 96

persistance d’une illusion 323
personnalité(s) 156, 165
de jugement 214, 376
et situations 215
philosophie 237
pièges de l’entretien d’embauche 313
plafond de prévisibilité 149
plans d’affaires (capital-risque) 169
poignée de main 316
police scientifique 13, 257
politologue 53
Portia 352, 354
pourcentage de concordance (PC) 112
pourcentage de déterminants communs 112
pouvoir discrétionnaire 19
précision 70
*Prédiction clinique et prédiction statistique : analyse théorique et examen des preuves* (Meehl) 118
prédiction(s) 111, 117
cliniques ou mécaniques 118, 318
*grade point average* ou GPA 123, 130
liberté sous caution 133
non régressives 188
par équivalence 186, 403
par vision externe 188
sans données historiques 131
prescription d’opioïdes 93
présomption irréfragable 354
pression sociale 106
prévention 255

des biais *ex ante* 248
prévision(s) 12, 32, 271
à long terme 147
calibrée 249, 276-277
de la performance 116
de la récidive 133
des parts de marché 60
probabilistes 275
Price, Mark 84
Princeton 92, 154-155, 157-159, 222
« principal » « agent » 363
*Principles of Forecasting* (Armstrong) 272
probabilité 49, 171
problèmes
difficiles 208
faciles 208
hypothétiques 52
réels 52
procès équitable 19, 338, 351
processus 53
de décision 335
de jugement 53
neuronaux endogènes 97
professeurs 55
propension 163
propositions de référendum 102
ProPublica 347
protection de l'enfance 12
protocole d'évaluations intermédiaires 14, 233, 324, 383

dans les décisions 333
prudence des experts 264
psychiatrie 295
psychologie 40
du bruit 165
Pythagore (théorème) 66, 80
QCM 371
question de la balle et de la batte 242
radiographie pulmonaire 290
radiologie 235, 288
Ramji-Nogales, Jaya 95
rapports d'intensités 202
réalisme naïf 34
recrutement 312
corrélation entre évaluations 315
intégration des données 318
premières impressions 315
réduction du biais 67, 247
*ex ante* 247
*ex post* 247
réestimation dialectique 89
référence (définition) 297
références 321
régimes 243
de retraite 248
règle(s)
contournement 356
règle(s) (directives) 355
complexe 123

principes généraux 359, 361

révisées 355

simples 117

régression

(linéaire) multiple 117, 128, 130, 284, 378

vers la moyenne 188

répétabilité intrajuges 291

ressemblance 171

ressources humaines 12

résultat(s) 53

et GMA 240

*Retour du Jedi, Le* (film) 174

rétroviseur 161

réunion 41, 104, 324, 327, 330, 384

Ritov, Ilana 244

robustesse 132

roman 31

*Roméo et Juliette* 175

Rosenzweig, Phil 227

rythme cardiaque 44

Salganik, Matthew 100-102, 154, 158

Salikhov, Marat 281

Satopää, Ville 280-283

Save More Tomorrow (plans) 248

Schkade, David 195

seconde loterie 85

seconde opinion 287

sélection 32, 271, 280

de candidats à l'embauche 144

Sellier, Anne-Laure 250

Sénat 23

sensibilité aux stéréotypes 91

*Sentencing Reform Act* (1984) 23

sentiment

de cohérence 52

de comprendre 163

séries 95

« Shared outrage and erratic awards » (Kahneman, Sunstein, Schkade) 195

signal intérieur 50, 52, 143-144, 378

« signature » d'un juge 216

Simkin, Amy 60-63, 67-69

Simonsohn, Uri 94

Slovic, Paul 175

smartphone 44

société de gestion d'actifs 32

sociologue 40

souscripteurs 28

souvenirs heureux, souvenirs malheureux 91

stabilité ou fidélité test-retest 86

stand de tir 9-10

*Star Wars* 174-175

statistiques 249

significatives 157

sténose du canal rachidien 289

stéréotype du leader 245

Stevens, S. S. 202-204

Stith, Kate 24, 338

stratégies de réduction du bruit

génératrices d’erreur 339

inconvénients 386

stress 93

structuration des jugements complexes 233, 278, 305, 312, 319, 383

style(s) cognitif(s) 235, 242

des meilleurs prévisionnistes 244

substitution 170

subtilité 123

Sunstein, Cass R. 195

supériorité

de la prédiction sur le jugement clinique 148

des règles et algorithmes 139

superprévisionnistes 148, 277

surprises 160

syndrome de la blouse blanche 286

Système 1 (pensée) 165, 167, 175, 186-188, 190, 208, 243, 404

Système 2 (pensée) 163, 175, 188, 243

*Système 1/Système 2, les deux vitesses de la pensée* (Kahneman) 167

systèmes 32

d’évaluation de la performance

réformes 303

taux de base 172, 278, 328

négligence du 173

téléchargements de chansons 100

témoins oculaires d’un crime 283

température 22, 94

tendon, dégénérescence 287

tension artérielle 286

test de dépistage rapide (TDR) 287

test de réflexion cognitive (*cognitive reflection test* ou CRT) 242
tests
de dépistage du cancer 291
standardisés 240
sur échantillons de travail 321
Tetlock, Philip 146-148, 274-277, 279-281
*The New Yorker* 364
Todorov, Alexander 222-223
traces digitales 259
traders 32
fraudes 28
trafiquant d'héroïne 21
train raté 70
traitement individualisé 351
trajectoires de vie des familles socialement fragiles 154
tramway
dilemme du 92
variante de la passerelle 92-93
travailleurs sociaux 221
trouble
anxieux 295
dépressif 296
tuberculose 290
Twitter 343
université
décisions d'admission 346
service des admissions 36
Université de Boston 249
Université de Pennsylvanie 96

Université du Minnesota 36, 118
Université hébraïque de Jérusalem 103
University College de Londres 260
utilitariste 92
vaccin 32
valeur prédictive 116
des informations 404
des tests 239
valeur(s)
et priorités, changement 353
valeur relative-valeur absolue 203
validité 264
variabilité
dans les pondérations 211
du médecin 296
interindividuelle 50
intra-individuelle 50
invisible 29
variété des compétences individuelles des juges 213
vérification 111
indépendante 262
vertus dissuasives 357
violence 365
visages 222
vision externe 162, 173, 187, 189, 309, 328, 355, 380, 382
Vucetich, Juan 258
Vul, Edward 87-90
Wainer, Howard N4
*Weapons of Math Destruction* (O'Neil) 346

Weber, Max 368-369

Welch, Jack 304

Williams, Thomas 21

*Woodson vs North Carolina* (cas) 344

*Work Rules !* (Bock) 319

Yale Law School 338

Yang, Crystal 25

Yu, Martin 125-126

Zuckerberg, Mark 241

# Remerciements

Les personnes que nous avons à remercier sont nombreuses. Linnea Gandhi a été notre chef de cabinet : outre son aide et ses conseils, c’est elle qui a organisé de bout en bout nos recherches et l’écriture de ce livre, tout en nous faisant rire et sourire. Elle nous a aussi apporté de précieuses suggestions sur le manuscrit. Rien n’aurait été possible sans elle. En coécrivant un des articles qui sont à la source de ce livre, Don Lovallo a joué un rôle majeur. John Brockman, notre agent, s’est montré positif, enthousiaste et avisé tout au long du projet. Tracy Behar, notre principale guide et relectrice, a permis d’améliorer de bien des façons ce livre. Arabella Pike et Ian Straus nous ont aussi apporté de précieux conseils éditoriaux.

Nous adressons des remerciements tout particuliers à Oren Bar-Gill, Maya Bar-Hillel, Max Bazerman, Tom Blaser, David Budescu, Jeremy Clifton, Anselm Dannecker, Vera Delaney, Itiel Dror, Angela Duckworth, Annie Duke, Dan Gilbert, Nathalie Gonzalez, Adam Grant, Anupam Jena, Louis Kaplow, Gary Klein, Jon Kleinberg, Nathan Kuncel, Kelly Leonard, Daniel Levin, Sara McLanahan, Barbara Mellers, Josh Miller, Sendhil Mullainathan, Scott Page, Eric Posner, Lucia Reisch, Matthew Salganik, Ville Satopää, Eldar Shafir, Tali Sharot, Philip Tetlock, Richard Thaler, Barbara Tversky, Peter Ubel, Duncan Watts, Caroline Webb et Crystal Yang, qui ont lu et commenté les premières versions des chapitres, et

quelquefois de l’ouvrage entier. Nous leur sommes très reconnaissants pour leur aide et leur générosité.

Nous avons eu la chance de bénéficier des conseils d’un grand nombre de chercheurs. Julian Parris nous a beaucoup aidés sur les questions statistiques. Nos chapitres sur l’apprentissage automatique n’auraient pas été possibles sans Sendhil Mullainathan, Jon Kleinberg, Jens Ludwig, Gregory Stoddard et Hye Chang. Et notre discussion sur la cohérence du jugement doit beaucoup à Alex Todorov et ses collègues de Princeton Joel Martinez, Brandon Labbree et Stefan Uddenerg, ainsi qu’à Scott Highhouse et Alison Broadfoot. Ces équipes de chercheurs exceptionnels n’ont pas seulement partagé gracieusement leurs idées : ils ont aussi eu la gentillesse de réaliser des études spécialement pour nous. Bien sûr, les malentendus ou les erreurs qui subsistent sont entièrement de notre responsabilité. Nous voudrions aussi remercier Laszlo Bock, Bo Cowgill, Jason Dana, Dan Goldstein, Harold Goldstein, Brian Hoffman, Alan Krueger, Michael Mauboussin, Emily Putnam-Hornstein, Charles Scherbaum, Anne-Laure Sellier et Yuichi Shoda, qui ont partagé avec nous leur expertise.

Nous sommes redevables à une véritable armée de chercheurs au fil des années, parmi lesquels Shreya Bhardwaj, Josie Fisher, Rohit Goyal, Nicole Grabel, Andrew Heinrich, Meghann Johnson, Sophie Mehtal, Eli Nachmany, William Ryan, Evelyn Shu, Matt Summers et Noram Ziv-Crispel. Une grande partie des discussions de ce livre impliquent des domaines dans lesquels nous manquons d’expertise, et grâce à leur excellent travail, ce livre est moins sujet au biais et au bruit qu’il ne l’aurait été sans eux.

Enfin, faire travailler ensemble trois auteurs sur deux continents n’est pas une tâche facile quand tout va pour le mieux, et en 2020, tout n’allait pas pour le mieux. Nous n’aurions pas terminé ce livre sans la magie technologique de Dropbox et de Zoom. Nous avons une pensée pour celles et ceux à qui nous devons ces outils.

# SOMMAIRE

INTRODUCTION - Deux types d'erreurs

PARTIE I - Détecter le bruit

- CHAPITRE 1 - Crime et châtiment
- CHAPITRE 2 - Un bruit systémique
- CHAPITRE 3 - Les décisions singulières

PARTIE II - Votre esprit est un instrument de mesure

- CHAPITRE 4 - Questions de jugement
- CHAPITRE 5 - Mesurer l'erreur
- CHAPITRE 6 - L'analyse du bruit
- CHAPITRE 7 - Le bruit occasionnel
- CHAPITRE 8 - Les groupes et le bruit

PARTIE III - Le bruit dans les jugements prédictifs

- CHAPITRE 9 - Jugements et modèles
- CHAPITRE 10 - Le silence des formules
- CHAPITRE 11 - L'ignorance objective
- CHAPITRE 12 - Dans la vallée des Choses Normales

PARTIE IV - D'où vient le bruit ?

- CHAPITRE 13 - Heuristiques, biais et bruit
- CHAPITRE 14 - Équivalences
- CHAPITRE 15 - Échelles

CHAPITRE 16 - Patterns

CHAPITRE 17 - Les composantes du bruit

PARTIE V - Améliorer les jugements

CHAPITRE 18 - De meilleurs juges pour de meilleurs jugements

CHAPITRE 19 - Réduction des biais et hygiène de la décision

CHAPITRE 20 - Police scientifique : le séquencement de l'information

CHAPITRE 21 - Prévision : l'agrégation et la sélection

CHAPITRE 22 - Médecine : les directives

CHAPITRE 23 - Évaluation de la performance : le choix de l'échelle

CHAPITRE 24 - Recrutement : la structuration des jugements complexes

CHAPITRE 25 - Le protocole d'évaluations intermédiaires

PARTIE VI - Le bruit optimal

CHAPITRE 26 - Les coûts de la réduction du bruit

CHAPITRE 27 - Une affaire de dignité

CHAPITRE 28 - Des règles ou des principes ?

SYNTHÈSE ET CONCLUSION - Prendre le bruit au sérieux

ÉPILOGUE - Moins de bruit

GLOSSAIRE - Jugement, erreur, biais, bruit

ANNEXE A - Conduire un audit de bruit

ANNEXE B - Observer une décision

ANNEXE C - Corriger les prédictions intuitives

Notes

Index

Remerciements

Des mêmes auteurs

# Des mêmes auteurs

## Daniel Kahneman

*Système 1, Système 2, les deux vitesses de la pensée*, Flammarion, 2012.

## Olivier Sibony

*Vous allez commettre une terrible erreur ! Combattre les biais cognitifs pour prendre de meilleures décisions*, Flammarion, 2019.

## Cass R. Sunstein

*Nudge. La méthode douce pour inspirer la bonne décision* (avec Richard H. Thaler), Vuibert, 2010.

Inscrivez-vous à notre newsletter !

Vous serez ainsi régulièrement informé(e)
de nos nouvelles parutions et de nos actualités :

https://www.odilejacob.fr/newsletter

www.odilejacob.fr

Suivez nous sur :

Odile Jacob

Dès qu'il y a jugement, il y a bruit. Quand deux médecins posent des diagnostics différents pour le même patient, quand deux juges attribuent des peines plus ou moins lourdes pour le même crime, quand deux responsables de ressources humaines prennent des décisions opposées à propos d'un candidat à un poste, nous sommes face au bruit.
Daniel Kahneman, Olivier Sibony et Cass R. Sunstein montrent dans ce livre que le bruit exerce des effets nocifs dans de nombreux domaines : médecine, justice, protection de l'enfance, prévision économique, recrutement, police scientifique, stratégie d'entreprise…
Pourtant, le bruit reste méconnu. Il est la face cachée de l'erreur de jugement. *Noise* nous propose des solutions simples et immédiatement opérationnelles pour réduire le bruit dans nos jugements et prendre de meilleures décisions.

**Le prochain livre qui va changer votre façon de penser.**
**Un best-seller mondial.**

« Un tour de force d'érudition et de clarté. »
THE NEW YORK TIMES

« Une leçon d'humilité devant nos erreurs. »
THE FINANCIAL TIMES

« Tous les chercheurs, les décideurs politiques, les dirigeants et les consultants devraient lire ce livre. »
THE WASHINGTON POST


DANIEL KAHNEMAN
est spécialiste de psychologie cognitive et d'économie comportementale, professeur émérite à Princeton et prix Nobel d'économie. Il est l'auteur de *Système 1/ Système 2. Les deux vitesses de la pensée.*

OLIVIER SIBONY
est professeur à HEC Paris et *associate fellow* de la Saïd Business School (Université d'Oxford), où il enseigne la stratégie et la prise de décision. Il est l'auteur de *Vous allez commettre une terrible erreur !*

CASS R. SUNSTEIN
est professeur de droit à Harvard. Il a également été directeur des affaires réglementaires au sein de l'administration Obama de 2009 à 2012. Il est l'auteur avec Richard Thaler de *Nudge. Comment inspirer la bonne décision.*

Design graphique de la couverture © Harper Collins Publishers 2021, images © Shutterstock

This file was downloaded from Z-Library project

*Your gateway to knowledge and culture. Accessible for everyone.*

z-library.sk z-lib.gs z-lib.fm go-to-library.sk

Official Telegram channel

Z-Access

https://wikipedia.org/wiki/Z-Library